I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 10545

True Overconfidence, Revealed through Actions: An Experiment

Stephen L. Cheung
Lachlan Johnstone

FEBRUARY 2017

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 10545

True Overconfidence, Revealed through Actions: An Experiment

Stephen L. Cheung
The University of Sydney and IZA

Lachlan Johnstone
The University of Sydney

FEBRUARY 2017

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

True Overconfidence, Revealed through Actions: An Experiment*

We report an experiment that infers true overconfidence in relative ability through actions, as opposed to reported beliefs. Subjects choose how to invest earnings from a skill task when the returns depend solely upon risk, or both risk and relative placement, enabling joint estimation of individual risk preferences and implied subjective beliefs of placing in the top half. We find evidence of aggregate overconfidence only in a treatment that receives minimal feedback on performance in a trial task. In treatments that receive more detailed feedback, aggregate overconfidence is not observed although identifiable segments of overand underconfident individuals persist.

JEL Classification: C91, D03, D81, D83

Keywords: true overconfidence, overplacement, subjective beliefs, joint estimation

Corresponding author:
Stephen L. Cheung
School of Economics
The University of Sydney
Merewether Building H04
Sydney NSW 2006
Australia

E-mail: Stephen.Cheung@sydney.edu.au

* We thank Kadir Atalay, David Butler, Pablo Guillen, Stephanie Heger, Lana Friesen, Guy Mayraz, Lionel Page, Curtis Price, Robert Slonim, Agnieszka Tymula, and participants in the Sydney Experimental Brownbag Seminar and Australia New ZealandWorkshop on Experimental Economics. This study was approved by The University of Sydney Human Research Ethics Committee, under protocol number 2016/427.

A well-known stylised fact in behavioural economics is that people appear to be overconfident. In arguably the most memorable demonstration of this idea, Svenson (1981) famously found that 88% of US respondents consider themselves to be safer than the median driver, while 93% rate themselves more skilful than the median. On the basis of evidence such as this, De Bondt and Thaler (1995, p. 389) assert that "(p)erhaps the most robust finding in the psychology of judgment is that people are overconfident". Taking this stylised fact as given, a lively literature in behavioural economics and finance proceeds to explore the implications of overconfidence for the decision making of consumers, investors, and managers.[1]

Recent literature distinguishes several forms of overconfidence (Moore and Healy, 2008). In this paper we focus on *overplacement*, or overconfidence in one's *relative* ability, as illustrated by the example above.[2] Economic experiments on overconfidence improve upon psychological studies such as Svenson (1981) by introducing financial incentives, and assessing an observable measure of skill (against an explicit reference population) within the experimental session itself. Contrary to the presumption of an overwhelming weight of evidence in psychology,[3] the case for overconfidence in experimental economics has in fact never been completely settled. We believe there are two related reasons why this is so.

First, the majority of studies (both in psychology and economics) infer overconfidence from reported beliefs, albeit that economists typically provide incentives for truthful reports. Yet this sits slightly uneasily with conventional revealed preference doctrine in economics, which holds that preferences (and by extension, beliefs) are revealed through actions not introspective statements. Belief elicitation mechanisms – such as the quadratic scoring rule (QSR), or an adaptation of the Becker et al. (1964, BDM) procedure to the elicitation of subjective probabilities – require a subject to consciously formulate a belief over (say) how likely she is to place in the top half, and report this to the experimenter. Yet no such process need occur when an individual behaves "as if" she holds an overconfident belief. Indeed, overconfidence has real consequences precisely when agents behave in an overconfident manner, as opposed to merely holding (or constructing) overconfident beliefs.

Second, until recently few studies collected sufficient information to refute the critique of Benoît and Dubra (2011), whereby the data could be generated by a population of rational Bayesians who have incomplete information regarding their abilities. Benoît and Dubra (2011) show how it is easy to construct examples in which more than 50% of a population can rationally believe they are more likely than not to be in the top half. For example, suppose it is common knowledge that a population is made up of good and bad drivers in equal proportion, that good drivers never have accidents, and 20% of bad drivers have accidents. Suppose the only signal a driver has of her type is whether or not she has an accident. Then 10% of drivers have accidents

[1]See Grubb (2015), Daniel and Hirshleifer (2015), and Malmendier and Tate (2015) for recent reviews of these respective bodies of research.

[2]Other forms of overconfidence include *overestimation* of one's absolute ability, and *overprecision* in the accuracy of beliefs. From here on, we use overconfidence to refer specifically to overplacement, unless otherwise noted.

[3]In fact, psychologists find that people may either be under- or overconfident depending on the difficulty of the task (Kruger, 1999; Moore and Cain, 2007), and have explored rational explanations for this finding (Kruger et al., 2008; Moore and Healy, 2008).

and know they are bad, while the other 90% know only that the Bayesian probability that they are good is $\frac{5}{9}$. If drivers were asked to report categorically whether they believe themselves to be good or bad, then 90% would claim to be good. But if instead they could be induced to truthfully report the probability with which they believe they are good, the average report would be $0.9\left(\frac{5}{9}\right) + 0.1\,(0) = 0.5$, which is the true population probability. Thus an experiment can only identify what Benoît and Dubra (2011) call "true overconfidence", inconsistent with any rationalising Bayesian model, by recovering these more detailed beliefs. Summarising the findings of the limited research that tests for true overconfidence, Benoît et al. (2015, p. 299) conclude that "a somewhat mixed picture emerges at this stage of research".

While some early economic experiments sought to reveal overconfidence through choice behaviour, they were not immune to the Bayesian critique. Subjects in Camerer and Lovallo (1999) choose whether to enter a market, where higher-ranked entrants receive a larger share of profits, and ranks are determined by performance in a test. In this setup the expected profitability of entry depends on the distribution of beliefs over all possible ranks, yet this distribution cannot be uncovered by the simple binary decision to enter or not. In Hoelzl and Rustichini (2005), subjects vote over whether prizes are awarded for placement in the top half in a test, or by a lottery with a 50% chance to win. From this, one can infer whether a subject's belief that she will place in the top half is greater or less than 50%, but not the exact strength of belief.[4]

Motivated in part by the Bayesian critique, most recent economic experiments revert to eliciting (incentivised) reports of belief. Here, the issues that arise mirror those encountered in a broader literature on eliciting subjective probabilities (over events not necessarily involving a subject's own skill), paramount being the confounding influence of a subject's risk preferences.[5] Broadly speaking, there are three main approaches. The first is to utilise a QSR, either with or without a control for risk preferences (see Clark and Friesen (2009), who control for risk preference using the binary lottery procedure of Berg et al. (1986); and Moore and Healy (2008), Merkle and Weber (2011), and Eil and Rao (2011), who do not make any risk correction). The second is to adapt the BDM mechanism for eliciting valuations to the elicitation of subjective probabilities (see Benoît et al. (2015), and Mobius et al. (2011)). The third approach, closest to the one we pursue in this paper, is to use a separate task to identify a subject's risk preferences, and use this to "correct" the inferred subjective beliefs (see Murad et al. (2016), who study overestimation of absolute ability; and Bruhin et al. (2016) who study how beliefs over relative ability affect risk taking, but do not test for true overconfidence).[6]

Andersen et al. (2014) are the first to jointly estimate risk preferences (utility and probability weighting parameters) and subjective beliefs, doing so in a context that does not involve judge-

[4]Moreover, both designs involve strategic interaction. Thus equilibrium predictions in Camerer and Lovallo (1999) assume risk neutrality, while the voting game in Hoelzl and Rustichini (2005) has multiple equilibria.

[5]See Schlag et al. (2015) for a review of the broader literature. Andersen et al. (2014), Harrison et al. (2014), and Offerman et al. (2009) describe specific techniques to control for risk preferences (including nonlinear probability weighting) in the QSR. See Karni (2009) for a statement of the BDM procedure for eliciting subjective probabilities. In what follows, we provide references specific to applications involving beliefs over own skill.

[6]A fourth approach, adopted by Burks et al. (2013), is to elicit the mode of the belief distribution by rewarding a subject for correctly predicting the quantile of her relative performance. Since this procedure elicits a different summary statistic of belief, it also necessitates a different approach to testing for overplacement.

ments over own skill.[7] They identify risk preferences using lottery choices involving known objective probabilities, and elicit reports of subjective beliefs using the QSR. As is well-known, the QSR is incentive compatible only under risk neutrality (Winkler and Murphy, 1970), even if this is not acknowledged by all who adopt it. In particular, if subjects are risk averse they will state reports that are less extreme than their true beliefs. Andersen et al. (2014) jointly estimate the preferences and beliefs that best explain observed responses in both tasks; their key finding is that estimated beliefs differ considerably from subjects' raw reports.

This paper reports an experiment that identifies true overconfidence in relative ability, robust to the Bayesian critique, revealed through choice behaviour instead of reported beliefs, and controlling for risk preferences through joint estimation of each subject's utility and probability weighting functions. We believe our work is unique in combining these features. As a further contribution, we study the effect of information upon the extent of overconfidence, by varying the level of feedback from a trial task on a between-subjects basis. We believe this to also be unique.[8] We find that as the quality of this signal becomes more precise, subjects hold better-calibrated beliefs in aggregate, although we also show that over- and underconfidence persist in the tails of the belief distribution. We also compare estimated to directly-reported beliefs; we find that they are highly-significantly correlated, and do not differ in levels. However, the estimated beliefs that best explain subjects' actions are more extreme (closer to 0 or 1) than their direct reports. Since our direct report data are unincentivised this finding is not a product of risk aversion, and may raise questions for the interpretation of direct reports.

Recent papers by Murad et al. (2016) and Bruhin et al. (2016) are methodologically closest to ours, although neither addresses our research question of testing for true overplacement. These studies similarly move away from reported beliefs, instead using binary choices over gambles in which the probability of payment depends upon performance in a test. Both papers also compare estimated beliefs to direct reports, albeit that Murad et al. (2016) do so (for beliefs over absolute performance) on a between-subjects basis, while Bruhin et al. (2016) apply the QSR *without* any correction for risk preference – despite the fact that their estimated beliefs incorporate just such a correction.

1 Experiment design

In our experiment, subjects make a series of allocation decisions over how to invest their earnings from a skill task. Payment for the experiment may thus depend on both performance in the task and the investment decisions. Subjects complete the main round of the skill task only after making their investment decisions. Thus the actual amount available to invest is not known at the time of the allocation decisions, which are framed in terms of shares of earnings invested.

[7] Their approach has antecedents in earlier work on time preference, most notably Andersen et al. (2008).

[8] In a distinct but complementary approach, Eil and Rao (2011) and Mobius et al. (2011) examine how subjects update their beliefs in response to feedback on performance, on a within-subjects basis. However, neither paper tests for true overconfidence.

In one set of decisions the return on investment depends only on risk, with the known probability of a high return varying over decisions. These choices enable us to identify each individual's risk preferences (utility and probability weighting functions). In a second set of decisions the return depends on both risk and whether the subject places in the top half of a group in the skill task. In conjunction with the first set, these choices enable us to jointly estimate each individual's implied subjective probability of placing in the top half. In a well-calibrated population, the mean belief should not deviate significantly from the true probability of one-half.

To examine the impact of signals of relative ability upon subjects' confidence, we vary the feedback that subjects receive on their performance in a prior trial of the skill task, on a between-subjects basis. For this feedback to be meaningful, it is essential that subjects are motivated to complete the trial to the best of their ability. Thus with one chance in three, a subject's payment for the experiment is determined by her performance in the trial.

1.1 Skill task

The skill task is adapted from Niederle and Vesterlund (2007). Within a time limit of five minutes, and with only working paper, subjects sum sets of five randomly-generated two-digit numbers to accrue earnings that increase linearly with each correct response. Each problem is presented on a new screen, and an unlimited number may be attempted within the time limit. Upon submitting an answer, the subject immediately moves on to the next screen, scoring one point if the response is correct but without receiving any immediate feedback.

Subjects complete two rounds of this task, each of five minutes duration: a trial prior to the two sets of investment decisions, and a main round afterwards. In the event that a subject is paid for the trial, she receives \$2.50 for each correct response.[9] Otherwise the subject accrues \$1.50 for each correct response in the main round, to be invested according to one of her allocation decisions, randomly selected and realised at the end of the session.[10]

1.2 Investment decisions: Risk

In the first set of decisions, which we refer to as risk decisions, subjects make nine allocations between a risk-free asset and a risky asset, framed neutrally as Accounts A and B. The risk-free asset yields a gross return of 1 with certainty. The risky asset yields a return of 3 with probability p or 0 otherwise, where p varies from 0.1 to 0.9 in increments of 0.1. Subjects allocate their earnings from the (subsequent) main round of the skill task by choosing a share s to invest

[9] All monetary values are expressed in Australian dollars. At the time of the experiments, 1 AUD was worth approximately 0.75 USD, 0.67 EUR, or 0.57 GBP.

[10] The reason for the higher piece rate in the trial is to help equalise expected payments between payment methods, given that the expected return on earnings invested from the main round is potentially greater than 1, while scores are also typically higher in the main round on account of learning.

in the risky asset, with each decision presented on a separate screen.[11] With one chance in three, a subject's payment for the experiment is determined by their earnings from the main round of the skill task, and the realisation of one randomly-selected decision from this set.

These decisions allow for a repeated measure of each subject's risk preference through exogenous variation in p. Each choice is in fact a modified version of the task in Gneezy and Potters (1997). The original version of their task suffices to identify a subject's degree of risk aversion under expected utility, but cannot discern risk-seeking from risk-neutral preferences. Repetition under varying probabilities allows a degree of risk seeking to be identified (since we include choices where the expected return on the risky asset is less than 1), as well as identifying the probability weighting function in models that allow for rank-dependent utility.

1.3 Investment decisions: Placement

In the second set of decisions, which we refer to as placement decisions, subjects make a further nine allocations. The same risk-free asset is available, however the return on the risky asset now depends both on chance and the subject's relative performance in the main round of the skill task, within an anonymous and randomly-assigned group of four. Should the subject place in the top half of this group, the risky asset yields a gross return of 3 with probability p or 1 otherwise; should she place in the bottom half, it yields 1 with probability p or 0 otherwise.[12] The assets are again framed neutrally as Accounts A and B, with each decision presented on a separate screen. As before, p varies from 0.1 to 0.9 in increments of 0.1, and subjects specify shares of earnings s to invest in the risky asset for each value. With one chance in three, a subject's payment for the experiment is determined by their earnings from the main round of the skill task, and the realisation of one randomly-selected decision from this set.

Whereas the optimal choices in risk decisions depend only on a subject's risk preferences, in placement decisions they depend also on her subjective belief regarding the probability of placing in the top half of the group (which we will denote by q). In the event of placing in the top half, the return on the risky asset is always at least as great as the risk-free return, irrespective of the value of p. Conversely, in the event of placing in the bottom half, it is the risk-free return that is always at least as great. Thus conditional upon risk preferences, as identified by the risk decisions, a subject who is quite confident of placing in the top half will invest in the risky asset even for low values of p, while one who is not-at-all confident will continue to prefer the risk-free asset even for high values of p.

[11] The share may be specified in any integer percentage. Subjects make decisions by placing or dragging a slider on a line with the mouse. As the slider moves, the screen updates to show the resultant distribution of returns from the chosen allocation for each dollar that the subject subsequently earns in the main round of the skill task. See the experiment instructions in Appendix C for a sample screen shot.

[12] Any ties within groups are resolved randomly to ensure that exactly two of the four group members receive the higher return on the risky asset. Subjects are informed of this procedure before making their decisions.

1.4 Information treatments

Immediately prior to making the placement decisions, each subject receives individual feedback on her performance in the trial round of the skill task, according to treatment. We vary the nature of this feedback on a between-subjects basis.

Subjects in the low information treatment (T1) are only told their raw score in the trial. They thus receive no information on their performance relative to others. Subjects in the medium information treatment (T2) are told their raw score as well as their ranking in the session of approximately twenty subjects. In the event of ties, subjects are told that they have tied and how many others have the same score as they do. Subjects in the high information treatment (T3) receive the most complete information. These subjects are told their raw score, their ranking in the overall session, and their ranking in the actual group of four to which they are assigned for the purpose of determining returns in the placement decisions that they will next make.

1.5 Direct statement of belief

After the placement decisions, and prior to the main round of the skill task, we ask subjects for a direct statement of belief. We express this in relative frequency terms, by asking the number of times out of 100 a subject believes she would place in the top half. This question appears without warning, and is posed at a moment when it is particularly salient for subjects to hold well-formed beliefs over their relative ability. While we do not suspect any motive for subjects to misrepresent these beliefs, since we do not incentivise the accuracy of their reports we do not rely upon them for our main results. Nonetheless they provide a useful consistency check and comparison to the implied beliefs that we infer from subjects' investment behaviour.

1.6 Procedures

The experiment took place at The University of Sydney over six sessions in August 2016. Each session ran for approximately 1.5 hours, and all subjects in a given session were assigned to the same information treatment. Between 16 and 24 subjects participated in each session, for a total of 40 in each treatment and an overall sample of 120. Subjects were recruited using ORSEE (Greiner, 2015), and the experiment was programmed in z-Tree (Fischbacher, 2007). In addition to receiving payment as determined by one of the three possible mechanisms outlined above, each subject also received a show-up fee of $5. Subjects were paid privately in cash as they left the laboratory, and the average payment inclusive of the show-up fee was $31.

Figure 1 summarises the experiment design. Appendix C provides the full text of the instructions for the low information treatment, and includes sample screen shots of the skill task and each of the two sets of investment decisions.

Figure 1: Summary of experiment design

1. Trial round of skill task; one-third chance to pay $2.50 per correct answer. *
2. First set of investment decisions (risk); one-third chance to pay for one decision. *
3. Feedback from trial round of skill task, according to treatment.
   - T1 (low information): raw score only.
   - T2 (medium information): as in T1, plus overall ranking within session of ~20 subjects.
   - T3 (high information): as in T2, plus ranking within assigned group of four.
4. Second set of investment decisions (placement); one-third chance to pay for one decision. *
5. (Unincentivised, surprise) direct elicitation of subjective belief.
6. Main round of skill task; $1.50 per correct answer, available to invest if paid according to 2. or 4.

Note: * denotes stages that only begin after all subjects correctly answer control questions.

2 Results

2.1 Basic descriptives

Table 1 presents summary statistics for the full sample, while Table A.1 in Appendix A reports Spearman rank correlation coefficients for the same variables (other than gender) in the full sample. Table A.2 in Appendix A reports the summary statistics disaggregated by treatment.

With the possible exceptions of payment, and the average share that a subject allocates to the risky asset in placement decisions, randomisation should ensure that the variables in Table 1 do not differ between treatments. We tested thoroughly for any differences, both in all treatments jointly and in each pairwise comparison between treatments (see Table A.3 in Appendix A for details). The only significant finding is that the treatments are not perfectly balanced with respect to age ($p = 0.004$, Kruskal-Wallis test). Since age does not correlate significantly with any of the other variables in Table A.1, we are confident that this does not drive our results.

The fact that scores in the main round of the skill task do not differ significantly between treatments ($p = 0.519$, Kruskal-Wallis test) suggests that our information treatments do not distort subjects' motivation to score highly in the task. Figure A.1 in Appendix A compares individual subjects' scores in the trial and main rounds of the task. While the difference between rounds is highly significant ($p < 0.001$, Wilcoxon signed-rank test, full sample), it takes the form of a uniform improvement across the skill distribution, with no individual subject differing conspicuously between the two rounds. This suggests that the difference is due to learning, as opposed to any strategic behaviour induced by our treatments.

Table 1: Summary statistics (full sample)

	Mean	SD	Median	10th pctl	90th pctl
Female	0.450				
Age	22.567	3.940	21.500	19.000	26.500
Payment	31.050	16.990	26.500	14.500	53.500
Skill task score (Trial round)	11.650	4.893	11.000	7.000	16.500
Skill task score (Main round)	13.092	5.102	12.000	9.000	18.000
Mean allocation: Risk	0.397	0.183	0.388	0.149	0.637
Mean allocation: Placement	0.455	0.341	0.433	0.000	0.999

Notes: Mean allocation refers to the mean share of experimental earnings that a subject allocates to the risky asset in, respectively, nine decisions in which the return on that asset depends only on risk, and nine decisions in which its return depends on both risk and the subject's placement in the top half of a group of four.

2.2 Investment behaviour

Figure 2 presents box plots of the share of earnings that subjects allocate to the risky asset in each risk and placement decision. In both sets of decisions, the amount invested in the risky asset increases monotonically as that asset becomes more attractive. The spread of allocations is consistently wider in placement decisions than in risk decisions, providing a first indication that subjects hold disparate beliefs over their likelihood of placing in the top half.

We find no significant difference between treatments in the mean amounts individual subjects allocate to the risky asset in risk decisions ($p = 0.616$, Kruskal-Wallis test). This suggests that the treatments do not systematically differ in risk preferences, as should be the case under random assignment. However the same test also finds no difference in the amounts allocated to the risky asset in placement decisions ($p = 0.211$), where feedback from the trial may have differential effects across treatments. Here, it is likely that a simple test based upon individual means may mask counteracting responses of subjects who receive good and bad news in higher information treatments. In the analysis of our structural estimates in Sections 2.5 and 2.6, we find a clearer picture of the effects of information on subjects' confidence.

The basic logic of our experiment design is to identify subjects' confidence of placing in the top half from the difference in behaviour in placement versus risk decisions, using the latter to control for risk preferences. As a simple model-free illustration of this idea, consider the difference between the mean amounts allocated to the risky asset in the nine placement decisions and the nine risk decisions. A subject for whom this measure is positive is willing to invest more when the return depends also on her relative performance, indicating confidence in placing in the top half of the group (and conversely for a subject for whom the measure is negative).

Figure 3 presents a jittered scatter plot of this difference, as a function of subjects' actual score in the main round of the skill task. The fitted line indicates a weak positive relationship: subjects who in fact score highly also tend to invest more in placement relative to risk decisions, consistent with greater confidence in their ability. However there is obviously considerable heterogeneity: subjects whose scores fall in the mass of the ability distribution appear to hold quite divergent beliefs. We next seek to quantify these differences more formally, in our structural estimates of subjects' risk preferences and subjective beliefs.

2.3 Structural estimation: Procedure

Our structural estimates are based upon a model of rank-dependent utility (RDU) (Quiggin, 1982), which nests the standard model of expected utility (EU) within it as a special case. Since all payoffs are in the gain domain, RDU also coincides with the cumulative prospect theory model of Tversky and Kahneman (1992). We assume that the utility function takes the constant relative risk aversion (CRRA) form:

$$u(x) = \begin{cases} \dfrac{x^{1-\theta}}{1-\theta} & \text{if } \theta \neq 1, \\ \ln x & \text{if } \theta = 1, \end{cases} \tag{1}$$

where, *under expected utility*, $\theta = 0$ corresponds to risk neutrality, $\theta > 0$ to risk aversion, and $\theta < 0$ to risk seeking.

This functional form implies that optimal allocations to the risky asset are invariant to the level of base earnings, which are as yet undetermined at the time subjects make their allocation decisions. These base earnings are constant over all decisions made by a given subject. Since we estimate models at an individual level, the CRRA specification fits naturally with our design.

Normalising base earnings to 1, in the nine risk decisions the rank-dependent utility of an allocation that invests a share s in the risky asset when the probability of a high return is p is:

$$RDU^R(s;p) = w(p)\,u(1+2s) + [1 - w(p)]\,u(1-s), \tag{2}$$

where $w(\cdot)$ is a probability weighting function, to be specified below.

Letting q denote a subject's belief regarding the probability with which she will place in the top half, in the nine placement decisions the rank-dependent utility of an allocation s is:

$$\begin{aligned} RDU^P(s;p,q) = \;& w(pq)\,u(1+2s) + [w(1-(1-p)(1-q)) - w(pq)]\,u(1) \\ & + [1 - w(1-(1-p)(1-q))]\,u(1-s). \end{aligned} \tag{3}$$

Since $1+2s \geq 1 \geq 1-s$ for all allocations $s \in [0,1]$, the decision weights can be written in this static manner: regardless of a subject's investment decision, the rank ordering over preferred states always remains the same.

Figure 2: Share of earnings allocated to the risky asset, by decision (full sample)

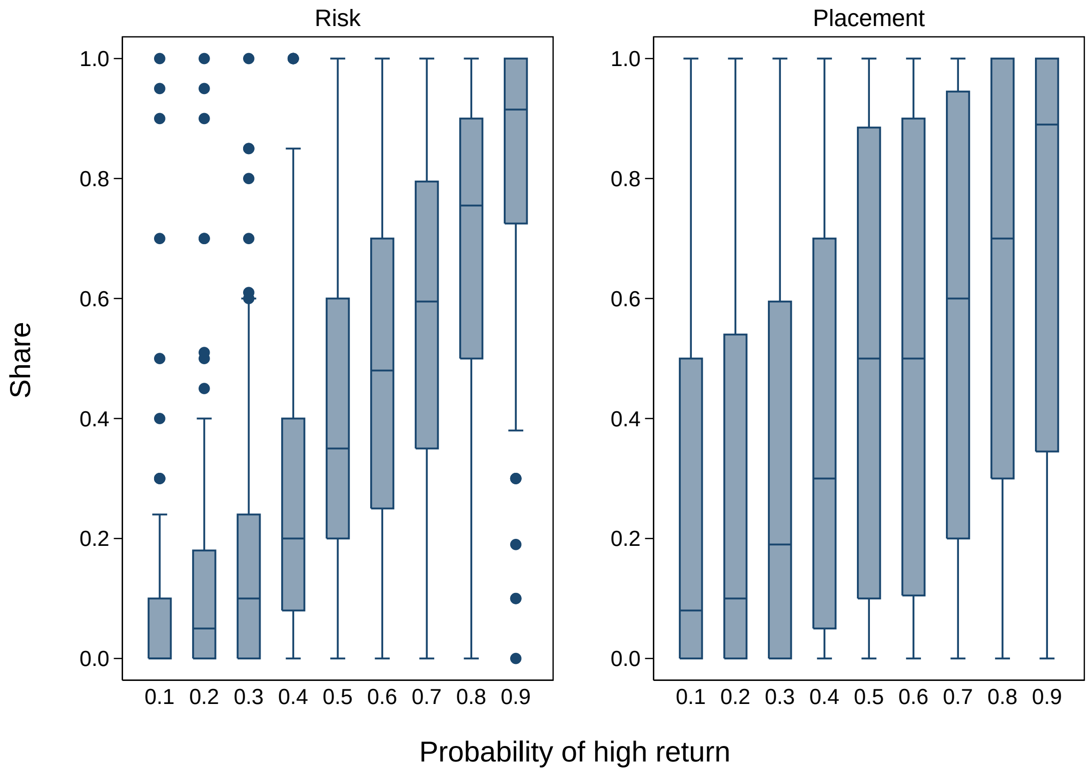
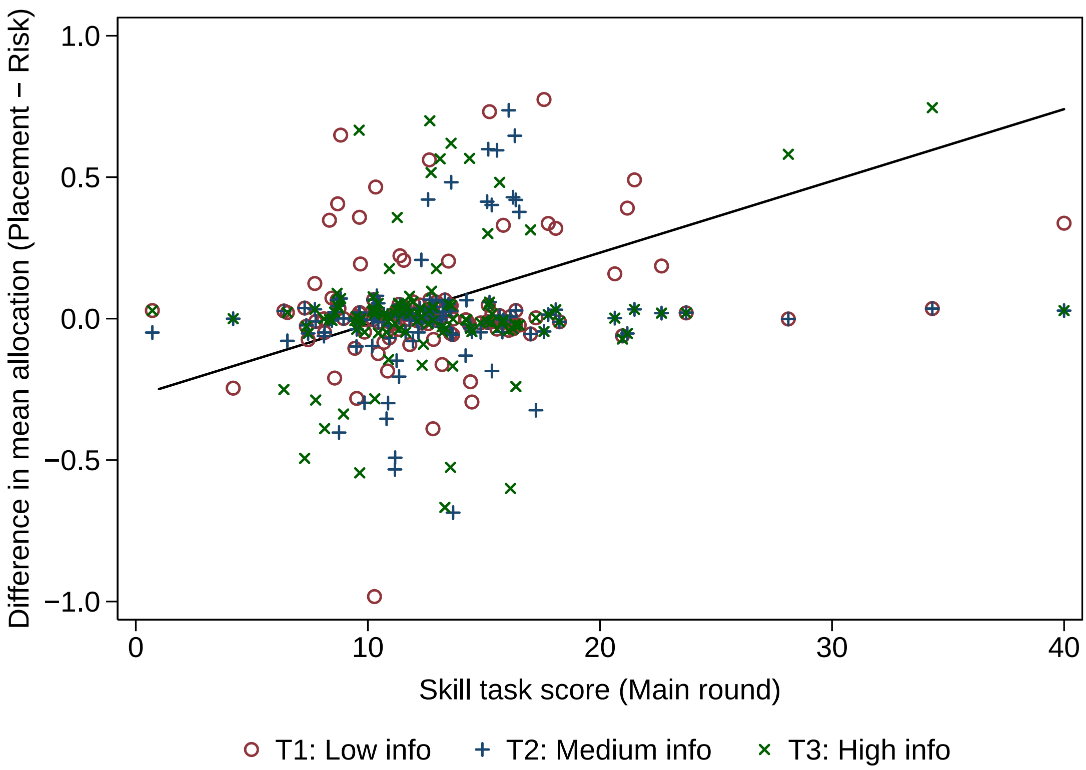


Figure 3: Differences in placement versus risk allocations, by treatment



We estimate models at an individual level for three specifications of the weighting function. The first is the standard model of expected utility, which imposes a linear weighting function:

$$w^{EU}(p) = p. \tag{4}$$

The second is a two-parameter function due to Goldstein and Einhorn (1987):

$$w^{GE}(p) = \frac{\delta p^{\gamma}}{\delta p^{\gamma} + (1-p)^{\gamma}}, \tag{5}$$

where $0 < \gamma < 1$ corresponds to an "inverse-S" shape, while increasing δ raises the point of intersection with the diagonal (Gonzalez and Wu, 1999). This specification reduces to a linear weighting function under the parameter restriction $\gamma = \delta = 1$.

Finally, we also estimate the two-parameter weighting function proposed by Prelec (1998):

$$w^{P2}(p) = \exp\left(-\beta\left(-\ln p\right)^{\alpha}\right), \tag{6}$$

where $0 < \alpha < 1$ corresponds to an "inverse-S" shape, while increasing β increases the convexity of the function. This specification reduces to a linear weighting function under the parameter restriction $\alpha = \beta = 1$.[13]

Harrison et al. (2013) provide compelling arguments for the use of a multinomial logit (MNL) estimation procedure in budget allocation designs such as ours – in preference to more conventional techniques such as nonlinear least squares (NLS) or Tobit[14] – and the MNL has been used by Harrison et al. (2013) and Cheung (2015) to model experimental data on time preference. Our preferred estimates are based on MNL, and we state the derivation of this estimator below. As a robustness check to our main estimates, we also report estimates obtained using NLS and Tobit; Appendix B derives these alternative estimation procedures.

The MNL recognises that the shares s allocated to the risky asset are discrete choices from a set of 101 possible alternatives, $S = \{0, 0.01, \ldots, 1\}$. In decision j subject i chooses $s_{ij} \in S$ to maximise the MNL utility specification:

$$RDU(s) + \varepsilon_{ijs}. \tag{7}$$

[13] Fehr-Duda and Epper (2012) recommend estimation of two-parameter weighting functions to capture heterogeneity in individual behaviour, and suggest that GE and P2 generally fit equally well.

[14] These arguments concern the manner in which the estimation procedures treat corner allocations, which are common in budget allocation designs (though less so in ones that involve risk – see Andreoni and Sprenger (2012) and Cheung (2015)). In brief, least squares suffers heteroskedasticity owing to the choice data being constrained between 0 and 1, while Tobit allows a latent variable to take values that are economically meaningless. By contrast, the MNL treats all allowable choices symmetrically such that corner allocations pose no specific concern.

The probability of the observed choice s_{ij} is thus:

$$\Pr\left(s_{ij}\right) = \Pr\left(RDU\left(s_{ij}\right) + \varepsilon_{ijs_{ij}} \geq RDU\left(s\right) + \varepsilon_{ijs}, \forall s \in S\right). \tag{8}$$

Under the error distribution in the MNL, this probability can be rewritten as (McFadden, 1974):

$$\Pr\left(s_{ij}\right) = \frac{\exp\left[\frac{1}{\lambda_i} RDU\left(s_{ij}\right)\right]}{\sum_{s\in S} \exp\left[\frac{1}{\lambda_i} RDU\left(s\right)\right]}, \tag{9}$$

where λ_i is the scale parameter in the distribution of subject i's utility error term.

The log-likelihood of a subject's observed choices in nine risk decisions ($j = 1, \ldots, 9$) is thus:

$$LL_i^R = \sum_{j=1}^{9} \ln\left(\frac{\exp\left[\frac{1}{\lambda_i} RDU^R\left(s_{ij}; p_j\right)\right]}{\sum_{s\in S} \exp\left[\frac{1}{\lambda_i} RDU^R\left(s; p_j\right)\right]}\right), \tag{10}$$

while that of the observed choices in the nine placement decisions ($j = 10, \ldots, 18$) is:

$$LL_i^P = \sum_{j=10}^{18} \ln\left(\frac{\exp\left[\frac{1}{\lambda_i} RDU^P\left(s_{ij}; p_j, q_i\right)\right]}{\sum_{s\in S} \exp\left[\frac{1}{\lambda_i} RDU^P\left(s; p_j, q_i\right)\right]}\right), \tag{11}$$

and the joint likelihood of the subject's choices in both sets of decisions is:

$$LL_i^{MNL} = LL_i^R + LL_i^P. \tag{12}$$

The maximum likelihood procedure finds estimates of utility curvature θ_i, probability weighting parameters (according to specification), a subject's implied subjective probability of placing in the top half q_i, and the scale parameter λ_i to maximise LL_i^{MNL} at an individual level. We estimate these parameters using optimisation routines in MATLAB version 2015b.[15]

[15] As the returns profile of assets in the risk decisions can only detect risk attitudes associated with a CRRA of no less than approximately -1 (under expected utility), the CRRA estimates are constrained to $\theta_i \geq -1.5$. Subjects with risk-seeking preferences beyond this would invest fully in the risky asset in every risk decision. Similarly, the maximum likelihood estimate of θ is capped at 30 for a subject who invests nothing in the risky asset in every risk decision. Finally, in specifications that allow for nonlinear probability weighting, we impose an upper bound of 10 on each of the weighting function parameters.

Table 2: Individual MNL parameter estimates (full sample)

	Mean	SD	Median	10th pctl	90th pctl
Expected utility					
θ: utility curvature	1.209	3.327	0.683	0.125	2.248
q: subjective belief	0.528	0.382	0.503	0.000	1.000
λ: scale	0.107	0.395	0.033	0.007	0.126
Rank-dependent utility (Goldstein-Einhorn)					
θ: utility curvature	1.450	4.665	0.197	0.000	2.862
γ: weighting, curvature	1.211	2.171	0.331	0.000	4.809
δ: weighting, elevation	1.639	2.711	0.527	0.418	3.914
q: subjective belief	0.544	0.377	0.532	0.000	1.000
λ: scale	0.046	0.350	0.003	0.000	0.029
Rank-dependent utility (Prelec)					
θ: utility curvature	1.713	4.336	0.208	0.000	4.666
α: weighting, curvature	1.644	2.784	0.393	0.000	6.675
β: weighting, elevation	1.476	1.648	1.099	0.369	2.619
q: subjective belief	0.533	0.384	0.503	0.000	1.000
λ: scale	0.045	0.318	0.001	0.000	0.030
Direct statement					
q: subjective belief	0.563	0.282	0.525	0.150	0.990

2.4 Structural estimation: Results

Table 2 summarises the individual MNL parameter estimates, for each specification of the probability weighting function, in the full sample. The final row summarises subjects' direct statements of q for comparison. Tables A.4, A.5 and A.6 in Appendix A report the same estimates disaggregated by treatment. Tables B.1 and B.2 in Appendix B summarise estimates obtained using the alternative NLS and Tobit estimation procedures. In each model, the median MNL estimate of θ corresponds to concave utility, while in specifications that allow for nonlinear probability weighting the median estimates correspond to an inverse-S shape. However there is considerable heterogeneity as evident from sizeable standard deviations.

To examine the precision of our MNL estimates, we construct confidence intervals using a parametric bootstrap procedure. Given the estimates for subject i, equation (9) specifies a probability mass function over all decisions j and budget shares s. We generate 999 bootstrap samples for each subject, each comprising nine risk and nine placement decisions for varying probabilities p, randomly drawn from this probability mass function. We obtain the bootstrap distribution of the estimates by re-estimating the model at each random sample, and apply the percentile method to this distribution to form 95% confidence intervals on the estimates.

Based upon this analysis, Table A.7 in Appendix A reports the number of subjects whose utility and probability weighting parameters differ from linearity at the 5% level. In all three models,

Table 3: Spearman rank correlation matrix of individual q estimates (full sample)

		MNL			*NLS*			*Tobit*			*Direct*
		EU	*GE*	*P2*	*EU*	*GE*	*P2*	*EU*	*GE*	*P2*	
	EU	1									
MNL	*GE*	**0.910**	1								
	P2	**0.875**	**0.954**	1							
	EU	*0.959*	0.930	0.889	1						
NLS	*GE*	0.881	*0.950*	0.927	**0.919**	1					
	P2	0.844	0.934	*0.936*	**0.877**	**0.966**	1				
	EU	*0.932*	0.840	0.805	*0.944*	0.837	0.767	1			
Tobit	*GE*	0.899	*0.937*	0.921	0.923	*0.960*	0.930	**0.866**	1		
	P2	0.852	0.917	*0.923*	0.877	0.947	*0.952*	**0.788**	**0.954**	1	
Direct		0.685	0.645	0.656	0.688	0.627	0.609	0.685	0.642	0.630	1

Notes: All correlations are significant at $p < 0.001$. Cells highlighted in bold compare estimates of alternative probability weighting specifications, holding the estimation technique constant. Cells highlighted in italics compare alternative estimation techniques, holding the probability weighting model constant.

over 75% of subjects have significantly concave utility. In both RDU specifications, subjects with inverse-S weighting functions are slightly outnumbered by those with the opposite pattern and pessimists outnumber optimists, although the respective parameters are significant for less than half of the sample. The final row reports subjects for whom both weighting parameters are jointly significant in Bonferroni simultaneous confidence intervals. These represent 33% of the sample in the GE specification, increasing to 50% in the P2 specification.

The primary data for our tests of overconfidence are the individual estimates of q, a subject's implied subjective probabilistic belief that she will place in the top half of the group, as identified by choices in placement decisions. Table 3 reports the Spearman rank correlation matrix of our estimates of q for alternative estimation techniques and specifications of the probability weighting function in the full sample. All correlations are highly significant at $p < 0.001$.

To facilitate interpretation of Table 3, cells that compare alternative weighting functions holding the estimation technique constant are highlighted in bold. These confirm that the RDU specifications consistently yield very similar estimates of q (Spearman $\rho \geq 0.954$); EU estimates differ somewhat but are nonetheless closely aligned with RDU ones ($\rho \geq 0.788$, with $\rho \geq 0.875$ in our preferred MNL estimates). Cells that compare estimation techniques holding the weighting function constant are highlighted in italics. These estimates are consistently tightly correlated ($\rho \geq 0.923$). Finally, the bottom row compares our estimates to subjects' (unincentivised) direct statements of belief. It is clear that these statements are somewhat distinct from beliefs inferred from actions ($\rho \in [0.609, 0.688]$), even though the correlations remain highly significant.

Table 3 indicates that our estimates robustly capture a stable primitive of behaviour that is related to, yet distinct from, stated beliefs. It makes little difference – for the rank-ordering of subjects with respect to implied confidence – which estimation technique is used, but it may make some difference whether nonlinear probability weighting is allowed for.

Figure 4 plots cumulative distribution functions (CDFs) for individual measures of q, disaggregated by treatment. The first three panels correspond to alternative probability weighting specifications in our preferred MNL estimates, while the final panel reports subjects' direct statements of q for comparison. Figure B.1 in Appendix B reports the corresponding CDFs for beliefs inferred using the alternative NLS and Tobit estimation techniques.

The CDF plots in Figure 4 highlight one reason why beliefs inferred from behaviour diverge somewhat from stated beliefs. Subjects are quite unlikely to self-report that they hold beliefs close to the extremes of 0 and 1, yet the structural estimates indicate that the behaviour of many can be best explained by just such beliefs. On the other hand subjects are more likely to directly state a belief of $q = 0.5$, and this is especially the case in the low information treatment.

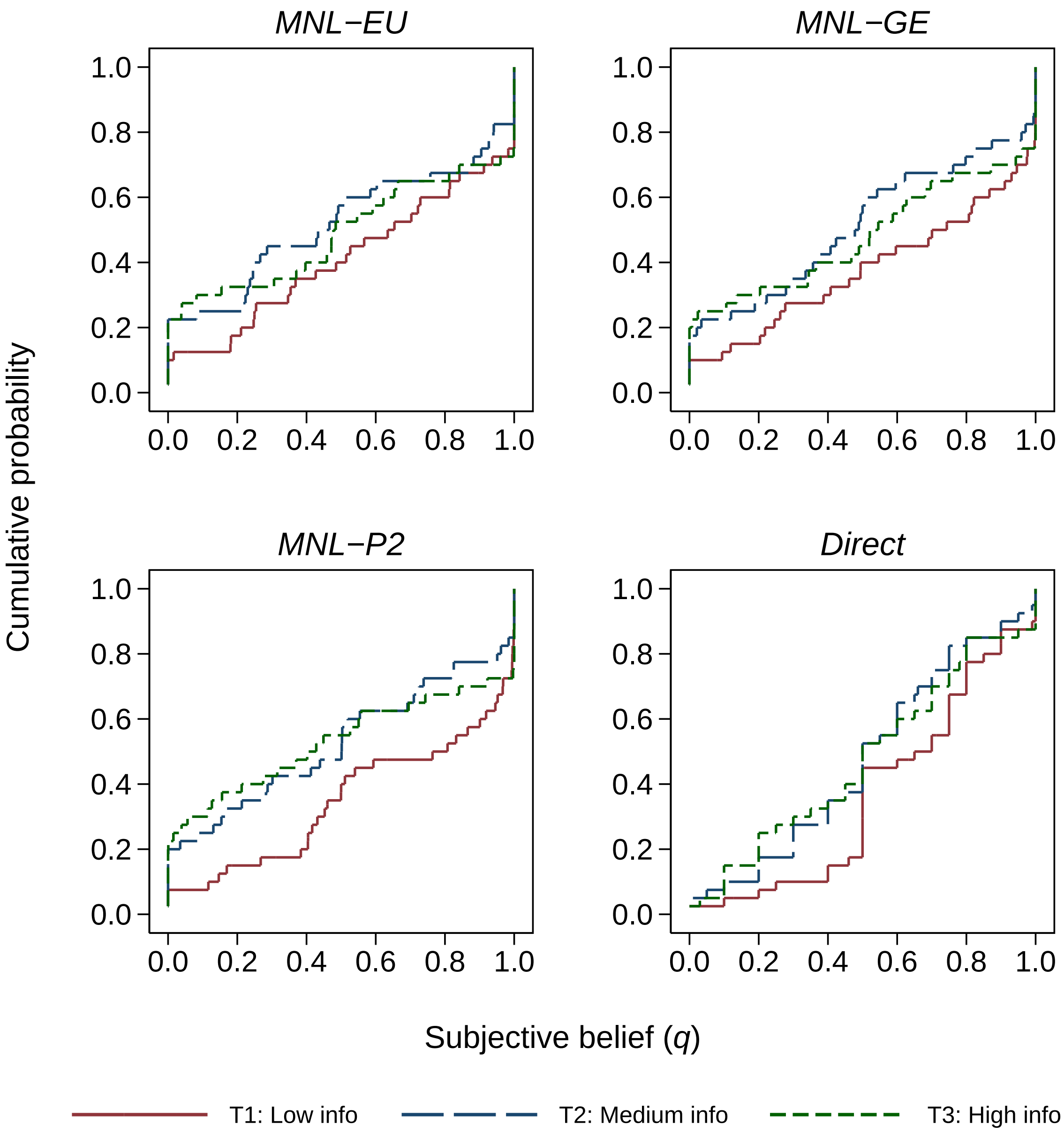
We examine the relation between estimated and reported beliefs in Table A.8 in Appendix A. In the top panel, we check for differences in levels in Wilcoxon signed-rank tests. For all three MNL models, and in both the full sample and each treatment, there is no significant difference in the central tendency of estimated and reported beliefs. Next, we examine the extremity of beliefs by collapsing them into three categories: $q \in [0, 0.1]$, $(0.1, 0.9)$, $[0.9, 1]$ and in the bottom panel test for differences in the distribution over these categories in Fisher exact tests. We find the difference to be highly significant in the full sample ($p \leq 0.001$), and at least marginally significant under high information ($p \leq 0.084$). In the remaining treatments we find at least a marginal difference in two of the three models.[16] This analysis confirms that subjects indeed report beliefs that are less extreme than those estimated from their behaviour.

Turning to treatment differences, Figure 4 generally indicates little systematic difference in the distribution of beliefs between the medium and high information treatments. However beliefs under low information appear more distinctive, with the CDF consistently located further to the right indicating that subjects in this treatment act as if they hold more confident beliefs. This is most evident in specifications that allow for nonlinear probability weighting. These observations also hold true for beliefs inferred using the two alternative estimation procedures (Figure B.1 in Appendix B). Moreover, the divergence between low information and the two higher information treatments can also be seen in subjects' direct statements of belief.

While Figure 4 provides preliminary indication that giving more detailed feedback on relative performance in the trial dampens subjects' confidence, it is unclear whether they are overconfident under low information, underconfident under medium and high information, or biased under all three. We next use our estimates to formally test whether subjects act upon beliefs that are well-calibrated when making their placement decisions, and how this varies with the

[16]We obtain even stronger results if we bound "extreme" beliefs more tightly by $q \leq 0.05$ and $q \geq 0.95$. In that case, the difference is significant at least at the 5% level in every test except for the EU model in the medium information treatment, where $p = 0.095$.

Figure 4: CDFs of MNL estimates of individual beliefs (q), by model and treatment



information provided in our treatments. For comparison, we will also report the same analyses using subjects' direct statements of belief.

2.5 Aggregate tests of overconfidence

A fundamental premise of the Bayesian view of overconfidence is that people are uncertain of their types: instead of believing with certitude that they either do or do not belong in the top half of the population, they continually update their beliefs in response to new information about their ability. The key implication is that "agents are not overconfident if they form their beliefs using the information available to them in an unbiased Bayesian manner" (Benoît et al., 2015, p. 302). Through our information treatments we can manipulate the signals of skill that subjects receive *within the course of the experimental session*, however we can neither observe nor control the signals they receive outside the laboratory. Since the experimenter cannot hope to observe all information used by a subject to derive her belief, it is all but impossible to identify a specific individual's belief as overconfident in this sense.

Nonetheless, Benoît and Dubra (2011) and Benoît et al. (2015) show that subjects' beliefs in aggregate must conform to certain minimal criteria to be consistent with any Bayesian model. In the following sections we test two such conditions, based on the theorems in Benoît et al. (2015). Firstly, at the level of the full sample and in each of the three treatments, we test whether the aggregate belief that subjects attach to placing in the top half matches the true probability of one-half. This first test uses information on every subject's (implicit) belief, but does not utilise any information on actual performance. Secondly, in Section 2.6, we identify sets of subjects in the upper and lower tails of the belief distribution whose actual performance (placement in the top half or not) is statistically incompatible with their beliefs. These tests enable us to identify over- and underconfident segments within a treatment, even where the beliefs of that treatment as a whole may appear to be well-calibrated.

Our aggregate tests of overconfidence are based upon Theorem 3 in Benoît et al. (2015). In our setting this states that the beliefs of a population, $\{q_i\}_{i=1}^{n}$, can be rationalised by a Bayesian model if and only if $\frac{1}{n}\sum_{i=1}^{n} q_i = 0.5$. Since the theorem is explicitly stated in terms of the mean belief, we follow Benoît et al. (2015) in reporting a parametric one-sample t-test. However our estimates of q are clearly non-normally distributed in higher information treatments, where they tend to be bimodal. For this reason we also report a nonparametric one-sample median test as a robustness check. While this does not assume normality it is not an exact test of the theory, which is expressed in terms of the mean. Fortuitously, we obtain a near-identical significance pattern using both tests, confirming that our results are indeed robust.

We report p-values for these tests in Table 4. We perform the tests both in the full sample, and separately in each information treatment. In the top panel, we use our preferred MNL estimates of subjects' beliefs as the data. The next two panels use estimates from the alternative NLS and Tobit procedures, and the final row uses subjects' direct statements of q for comparison.

Table 4: Tests of the hypothesis that $q = 0.5$ (two-sided p-values)

	One-sample *t*-tests				One-sample median tests			
	Full	T1: Low	T2: Med	T3: High	Full	T1: Low	T2: Med	T3: High
				Multinomial logit (MNL)				
EU	0.421	0.114	0.766	0.863	0.310	0.088 *	0.707	0.610
GE	0.205	0.036 **	0.820	0.747	0.267	0.046 **	0.657	0.697
P2	0.346	0.007 ***	0.677	0.645	0.455	0.020 **	0.610	0.727
				Nonlinear least squares (NLS)				
EU	0.062 *	0.014 **	0.943	0.427	0.041 **	0.013 **	0.872	0.166
GE	0.344	0.031 **	0.699	0.971	0.365	0.027 **	0.747	0.989
P2	0.455	0.014 **	0.504	0.737	0.471	0.019 **	0.493	0.819
				Tobit				
EU	0.001 ***	0.001 ***	0.235	0.203	0.003 ***	0.002 ***	0.301	0.327
GE	0.456	0.020 **	0.436	0.871	0.488	0.022 **	0.444	0.893
P2	0.570	0.029 **	0.454	0.799	0.441	0.027 **	0.554	0.968
				Direct statement				
Direct	0.016 **	0.002 ***	0.545	0.588	0.014 **	0.003 ***	0.540	0.558

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Using MNL estimates of subjects' beliefs, we find no evidence of aggregate overconfidence in the full sample, or the medium or high information treatments. However we do see evidence of overconfidence in the low information treatment. This emerges most clearly in specifications that allow for nonlinear probability weighting. As we will see in Section 2.6, this may be related to the fact that the MNL-EU specification also identifies a sizeable set of underconfident subjects in the low information treatment, whereas the two RDU models do not.

Using the alternative NLS and Tobit procedures to estimate subjects' beliefs, we confirm no aggregate overconfidence under medium or high information. Moreover these models consistently find overconfidence under low information, even in an EU specification (where they also find some evidence of overconfidence in the full sample). Finally, it is encouraging that inferences from subjects' direct statements of belief align well with those of our structural estimates: the direct reports indicate aggregate overconfidence under low information (as well as in the full sample), but not under medium or high information.

Overall, Table 4 indicates probable evidence of aggregate overconfidence under low information, but not under medium or high information. Insofar as there is overconfidence in the full sample, it appears to be driven by the low information group. However there remains the possibility that, in treatments where we do not detect aggregate overconfidence, the underconfident beliefs of some may cancel out the overconfident beliefs of others. We examine this possibility next, by identifying segments in the tails of the belief distribution whose actual placement in the skill task is inconsistent with their beliefs.

2.6 Tests of over- and underconfidence in segments of the belief distribution

Our "segment" tests of over- and underconfidence in the upper and lower tails of the belief distribution are based upon Theorem 2 in Benoît et al. (2015). In our setting, this states that if a fraction x of the population believe they are in the top half with probability at least $\bar{q}$ (where $\bar{q} > 0.5$), these beliefs can be rationalised if and only if $x\bar{q} \leq \tilde{x}$, where $\tilde{x}$ is the fraction who both hold those beliefs and who are indeed in the top half.

For example, if 40% of a population believe with $q_i \geq \bar{q} = 0.8$ that they are in the top half, the fraction who both hold this belief and place in the top half should be at least $\tilde{x} \geq 0.4 \times 0.8 = 0.32$. Equivalently, the likelihood of placing in the top half, conditional upon holding the belief $q_i \geq 0.8$ should be at least $\tilde{x}/x \geq \bar{q} = 0.8$. In short, absent the randomness of small samples, the probability that members of this segment ascribe to being in the top half must be correct.

The figures in the above example correspond approximately to the beliefs of the upper tail of the low information treatment as estimated by our MNL-EU model (see Table 5, top left cell). The proportion of this segment who in fact place in the top half is only $9/16 \approx 0.56 < 0.8$, and so our data fail the test described above. However, this data comes from a finite sample (of size $n = 40$), whereas the underlying theorem only holds true in a large population.

Benoît et al. (2015, Appendix C) extend their theorem to show that in a rational population, the probability of drawing a sample of size n that is at least as apparently overconfident as the data is $\Pr(b \leq n\tilde{x})$, where b is a random variable with binomial distribution $b \sim B(nx, \bar{q})$.

Continuing the example, we have a sample of $n = 40$, of whom a fraction $x = 16/40 = 0.4$ hold the belief $q_i \geq 0.8$, while a fraction $\tilde{x} = 9/40 = 0.225$ both hold this belief and in fact place in the top half. The probability that data at least as overconfident as this would arise from a rational population is $\Pr(b \leq 9) = 0.027$ for $b \sim B(16, 0.8)$, the cumulative binomial distribution with 16 trials, 9 successes and probability of success 0.8. (As seen in the top left cell of Table 5, for our low information treatment and MNL-EU estimates the actual cutoff for the beliefs of the segment we identify as overconfident is $q_i \geq 0.812$, therefore we are able to reject the null of rationality at a slightly higher significance of 0.019.)

Analogous arguments can be used to construct a test of underconfidence in the lower tail of the belief distribution. For example, Table 5 also shows that in the low information treatment and using our MNL-EU estimates, we identify a set of 11 subjects who believe with $q_i \leq 0.254$ that they will place in the top half. Yet 6 of this 11 in fact place in the top half. Equivalently, these subjects believe with probability $r_i \geq 0.746$ that they will place in the *bottom* half, when in fact only $5/11 \approx 0.454 < 0.746$ do so. The probability that a rational population would generate data at least as *underconfident* as this is $\Pr(b \leq 5) = 0.037$ for $b \sim B(11, 0.746)$.

Table 5 characterises the sets of subjects whom we identify as over- and underconfident according to these tests. For each estimation technique and weighting function, we rank order subjects within treatments from most to least confident according to the estimates of q_i from that model. We then identify the overconfident segment within a treatment as the largest contiguous set of subjects in the upper tail of the belief distribution, whose actual placement in the top half is inconsistent with the null of rational beliefs at a significance level of $p < 0.05$; we define the underconfident segment (if any) analogously in the lower tail. Within each cell of Table 5, the top entry shows the number of subjects who make up the over- or underconfident segment and the number of these who in fact place in the top half of their groups, while the bottom entry (in parentheses) shows the cutoff value of q_i for that set of subjects.[17]

Table 5 shows that all models consistently identify a sizeable overconfident segment in the low information treatment. Around 40 to 50% of subjects in this treatment behave as if they believe they have at least a 70 to 80% chance (according to specification) to place in the top half, whereas the fraction who do so is considerably lower than such beliefs would suggest. On the other hand, the MNL-EU model alone identifies any meaningful underconfidence under low information. As noted earlier, this may help explain why that model finds at best weak evidence of aggregate overconfidence under low information in Table 4: the biases of the over- and underconfident segments cancel out in aggregate tests. Nonetheless, given that the underconfident segment largely disappears in the more flexible RDU models, our results indicate systematic overconfidence under low information.

Turning to medium and high information, recall that the tests in Table 4 consistently find no aggregate overconfidence in these treatments. The more disaggregated tests in Table 5 suggest that there are in fact both over- and underconfident segments simultaneously present. In these treatments, feedback on performance in the trial will tend to lead subjects to act upon more strongly bifurcated beliefs. If many subjects hold the belief that they will each place in the top half with probability close to one, it suffices that only a small number fail to do so for a segment to be classified as overconfident (and analogously in the lower tail). Thus the behavioural response to information itself likely contributes to the more extreme beliefs identified in the over- and underconfident segments of the medium and high information treatments.

The tests in Table 5 are based upon subjects' placement in the top half within a randomly-selected group of four subjects from the same session. This is the outcome that is relevant to their earnings in placement decisions, and thus the one over which they need to form accurate beliefs for the purpose of making optimal choices. Moreover, through our use of a random tie-breaking rule – which is known to the subjects – it occurs with an aggregate probability of exactly one-half.[18] Nonetheless, these same procedures also introduce random noise into the

[17] We also perform the same tests using subjects' direct statements of belief (bottom row of Table 5). Because subjects are quite unlikely to directly assert beliefs close to 0 or 1 – even though the actions of many are best explained by just such beliefs – these tests find only limited violations of the null of rationality: a set of five overconfident subjects under low information, three under medium information, and no underconfident segments.

[18] There are eight subjects who miss out on a place in the top half on account of the random tie-break: three under low information, two under medium information, and three under high information. Thus $68/120 = 56.7\%$ of subjects perform well enough to potentially warrant a place in the top half within their group of four.

Table 5: Segment tests of over- and underconfidence

		T1: Low info		T2: Medium info		T3: High info	
		Over	Under	Over	Under	Over	Under
MNL	*EU*	16/9	11/6	7/6	10/3	12/9	11/3
		(0.812)	(0.254)	(1.000)	(0.082)	(0.960)	(0.040)
	GE	20/11	4/1	9/7	9/2	12/9	11/4
		(0.744)	(0.000)	(0.959)	(0.034)	(0.943)	(0.106)
	P2	20/12	0	6/5	8/1	11/9	12/4
		(0.808)	(n/a)	(1.000)	(0.001)	(0.997)	(0.056)
NLS	*EU*	20/11	0	9/6	0	5/4	11/3
		(0.745)	(n/a)	(0.921)	(n/a)	(0.993)	(0.077)
	GE	18/11	0	7/5	7/2	6/5	12/4
		(0.820)	(n/a)	(0.947)	(0.017)	(0.994)	(0.121)
	P2	19/11	0	7/5	12/4	9/7	13/5
		(0.796)	(n/a)	(0.964)	(0.099)	(0.983)	(0.147)
Tobit	*EU*	21/10	4/2	11/8	7/1	11/9	8/2
		(0.698)	(0.095)	(0.925)	(0.001)	(0.972)	(0.008)
	GE	20/11	3/1	4/3	9/3	8/7	12/5
		(0.770)	(0.001)	(0.995)	(0.035)	(0.997)	(0.143)
	P2	17/10	0	8/6	11/4	9/8	13/5
		(0.791)	(n/a)	(0.971)	(0.130)	(0.995)	(0.163)
Direct		5/4	0	3/2	0	0	0
		(0.990)	(n/a)	(0.990)	(n/a)	(n/a)	(n/a)

Notes: This table identifies sets of subjects in the upper and lower tails of the belief distribution whose actual placement in the skill task is not consistent with their beliefs at a significance of $p < 0.05$. In each cell the top entry shows the number of subjects who make up the over- or underconfident segment and the number of these who in fact place in the top half. The bottom entry (in parentheses) shows the cutoff value of q_i for subjects who make up this segment. For example, the *MNL-EU* model identifies an overconfident segment of 16 subjects in the low information treatment, who hold beliefs $q_i \geq 0.812$. Of these, 9 place in the top half. The chance that a sample at least as apparently overconfident as this would arise from a rational population is $\Pr(b \leq 9) = 0.019$, for $b \sim B(16, 0.812)$.

relationship between a subject's performance and her placement in the top half: a subject may in fact place in the top half (of the full sample, treatment, session, or even – allowing for ties – the group of four), and yet not be rewarded accordingly.

In Table 6 we examine the robustness of the segment tests to alternative definitions of placement in the top half, by relaxing the random tie-break and varying the comparison group used to evaluate relative placement, using our preferred MNL estimates as the data. In each case, we handle ties by adopting the definition that provides for the most stringent test. Thus, when a subject ties for a place in the top half, we consider that subject to belong in the top half when testing for overconfidence, but the bottom half when testing for underconfidence. The top panel continues to define placement within the randomly-selected group of four. The middle panel defines placement relative to the session, given that subjects in the medium information treatment receive a noisy signal of their ability relative to this group in feedback from the trial. Finally, the bottom panel defines placement relative to the full sample of 120 subjects.[19]

Table 6 confirms that the finding of an overconfident segment under low information is quite robust, whereas the underconfident segment which was observed in the EU specification largely disappears once the random tie-break is removed. This strengthens our interpretation that the low information group are indeed overconfident. Under high information the results point to roughly equal-sized over- and underconfident segments, each accounting for roughly one-quarter of the subjects in this treatment. Finally, the results for medium information are mixed; however if we consider the session to be the most relevant comparison group then subjects in this treatment appear remarkably close to being well-calibrated.

3 Discussion

This paper reports an experiment to infer true overconfidence, immune to the Bayesian critique of Benoît and Dubra (2011), through actions as opposed to reported beliefs. This approach avoids the need for elaborate mechanisms to incentivise truthful reports of belief, and sits more comfortably with revealed preference doctrine in economics. Our design permits control for risk preference, including nonlinear probability weighting, through joint estimation of individual subjects' risk preferences and subjective beliefs. We use these estimates to test for overconfidence both in aggregate and in the tails of the belief distribution, identifying under- as well as overconfident segments. In our information treatments, we vary the feedback that subjects receive on their relative performance in a trial task. We find aggregate overconfidence only in the treatment that receives the least feedback. Even in treatments in which aggregate overconfidence is absent, we find identifiable segments of over- and underconfident individuals. We also compare our estimated beliefs to subjects' (unincentivised) direct reports. While they are highly-significantly correlated and do not differ in levels, we find that subjects tend to report beliefs that are less extreme than the estimated beliefs that best explain their actions.

[19] In the full sample the median subject correctly solves 12 addition problems within the five-minute limit, with 18 subjects (15%) tied on this score, 55 (46%) who score better, and 47 (39%) worse. We thus treat 61% of subjects as placing in the "top half" when testing for overconfidence, but only 46% when testing for underconfidence.

Table 6: Segment tests of over- and underconfidence: Robustness checks

		T1: Low info		T2: Medium info		T3: High info	
		Over	Under	Over	Under	Over	Under
Rank in group, No tie break							
	EU	13/8	0	7/6	9/1	11/10	9/2
		(0.913)	(n/a)	(1.000)	(0.000)	(0.998)	(0.000)
MNL	*GE*	15/10	0	9/7	0	0	10/2
		(0.911)	(n/a)	(0.959)	(n/a)	(n/a)	(0.024)
	P2	18/12	0	6/5	0	11/10	12/3
		(0.866)	(n/a)	(1.000)	(n/a)	(0.997)	(0.056)
Rank in session, No tie break							
	EU	13/8	4/1	0	9/1	11/9	11/3
		(0.913)	(0.000)	(n/a)	(0.000)	(0.998)	(0.040)
MNL	*GE*	15/10	0	0	0	10/9	11/4
		(0.911)	(n/a)	(n/a)	(n/a)	(0.998)	(0.106)
	P2	18/12	0	0	0	11/10	12/4
		(0.866)	(n/a)	(n/a)	(n/a)	(0.997)	(0.056)
Rank in full sample, No tie break							
	EU	13/8	4/1	0	10/3	11/9	11/3
		(0.913)	(0.000)	(n/a)	(0.082)	(0.998)	(0.040)
MNL	*GE*	19/11	0	0	9/2	10/9	11/4
		(0.807)	(n/a)	(n/a)	(0.034)	(0.998)	(0.106)
	P2	20/12	0	0	8/1	11/10	12/4
		(0.808)	(n/a)	(n/a)	(0.001)	(0.997)	(0.056)

Notes: This table reports robustness checks of the segment tests in Table 5, using alternate definitions of placement in the top half. To provide for the most stringent tests, when a subject ties for a place in the top half within the relevant comparison group, we consider that subject to belong in the top half when testing for overconfidence, but the bottom half when testing for underconfidence. Sets of subjects are classified as over- or underconfident if their actual placement is not consistent with their beliefs at a significance of $p < 0.05$. In each cell the top entry shows the number of subjects who make up the over- or underconfident segment and the number of these who in fact place in the top half, as defined above, within the relevant comparison group. The bottom entry (in parentheses) shows the cutoff value of q_i for subjects who make up this segment.

Our multinomial logit estimates of q, the subjective probability of placing in the top half, represent the beliefs that best explain – in a maximum likelihood sense – the choice behaviour of each subject in placement decisions, assuming her objective is to maximise rank-dependent utility. While RDU allows for nonlinear probability weighting, it does not capture a variety of other considerations that may potentially enter the decision-making process.

First, the choices that a subject faces in placement decisions are inherently ambiguous, since they depend not only on her own skill but also that of unknown others. If the subject is ambiguity averse, her beliefs can no longer be expressed as a point probability. The multiple priors maxmin expected utility model (Gilboa and Schmeidler, 1989) proposes that the subject evaluates her utility under each probability distribution that she considers possible, and makes the choice that maximises the minimum of these utilities, in effect acting upon her most pessimistic belief. In this case, our estimates would represent a lower bound on subjects' beliefs, potentially understating the extent of overconfidence.

Second, subjects' beliefs may be biased by optimism (Weinstein, 1980), where the distinction between optimism and overconfidence is that optimism (or "wishful thinking") overstates the probability of a desired state independent of the subject's own performance. Heger and Papageorge (2016) argue that many scenarios in which overconfidence is studied also contain the possibility for optimism. The RDU model allows a form of optimism through overweighting the probability of high payoff states – *even with known objective probabilities in risk decisions* – if the weighting function is concave or inverse-S shaped. However if a subject acts upon unrealistic beliefs in placement decisions, whether for reasons of optimism or overconfidence, this is captured in our estimate of q. Heger and Papageorge (2016) design an experiment to disentangle optimism from overestimation of absolute performance. They find that the biases are positively correlated at an individual level, such that overlooking the former will overstate the latter. In our context of relative overplacement, optimism could have a similar effect.

Next, whereas ambiguity aversion and optimism operate through beliefs, there are also reasons why a subject may have a direct preference to invest in her own performance. The subject may prefer to take control of her own destiny (Heath and Tversky, 1991; Goodie, 2003), making investment in the risky asset more attractive in placement compared to risk decisions, or may have a taste for competition (Niederle and Vesterlund, 2007).[20] Finally, if the subject anticipates future self-control problems, she may invest in her own performance as a commitment device to motivate herself to work harder in the subsequent skill task (Kaur et al., 2015).

We cannot rule out the possibility that one or more of these forces are at work in our experiment, causing our estimates of q to be biased. However none provides a natural explanation

[20] In contrast to Niederle and Vesterlund (2007), whose skill task we adopt, we do not pursue the issue of gender differences in overconfidence. The reason we hesitate to do so is that, on our interpretation of the Bayesian critique of Benoît and Dubra (2011), it is unclear whether such differences can be attributed to true overconfidence unless it is assumed that males and females do not differ systematically in the signals of skill they receive outside the laboratory. While this holds true for our treatment groups by randomisation, it need not hold when the sample is partitioned on any other characteristic. In our data, males indeed act more confidently than females, even though males and females are equally likely to place in the top half. Nonetheless it is possible that the beliefs of both groups could be rational if males and females receive different signals outside the experiment.

for our main comparative static finding of reduced overconfidence in higher information treatments. More information should reduce the ambiguity subjects face, leading them to act upon less pessimistic worst-case beliefs, and appear more overconfident on average. If optimism is an unrealistic belief independent of performance, it should not be influenced by feedback on performance. Finally, the effect of information is replicated in subjects' (unincentivised) direct reports, where the preference-based explanations have no force.

While our finding regarding the effect of information sits comfortably with a Bayesian interpretation, it has not been seen in previous studies. Benoît et al. (2015) manipulate whether subjects are shown the population distribution of test scores on a between-subjects basis, and find no effect. In within-subject studies of belief updating, Eil and Rao (2011) and Mobius et al. (2011) find that subjects process positive and negative feedback asymmetrically, updating more fully in response to good than to bad news. If the subjects who receive good and bad news in our higher information treatments were to respond in the same manner then overconfidence would *increase* with information, but this is not what we find.

Despite similar implications in aggregate, there are qualitative differences between estimated beliefs and subjects' direct reports:[21] subjects behave as if they hold more extreme beliefs than the ones they explicitly report to the experimenter. Of course, there are reasons to be sceptical of our direct reports given that they are unincentivised. For example, subjects may report beliefs that rationalise their choices in placement decisions, causing the similarity between reported and estimated beliefs to be overstated. On the other hand, unincentivised reports at least have the advantage that they are not confounded by risk preferences: while risk aversion might similarly induce subjects to report less extreme beliefs under a QSR, this cannot account for the effect we observe in unincentivised reports. Finally, the format in which probabilities are represented may matter. Price (1998) finds that beliefs over absolute performance are less extreme when elicited as relative frequencies as opposed to probabilities. Since we express our direct report in relative frequency terms, this may contribute to the difference we observe.

Moreover, the use of incentivised reports brings trade-offs of its own. Both the QSR and the BDM mechanism (upon which the procedures in Benoît et al. (2015) and Mobius et al. (2011) are based) are cognitively challenging for subjects. Further, if subjects are paid for both their reported belief and performance in the skill task, they have a stake in the outcome they are asked to predict. In this case, standard belief elicitation techniques fail (Kadane and Winkler, 1988; Karni and Safra, 1995).[22] Our contribution in this paper is to show how an alternative approach of inferring beliefs from revealed choice behaviour can yield equally rich data – sufficient to test for true overconfidence – without encountering any of these difficulties.

[21] Murad et al. (2016) find that estimated beliefs and (unincentivised) direct reports yield distinct patterns of absolute overestimation in a between-subjects comparison, while Bruhin et al. (2016) find only modest correlation between estimated beliefs and reports incentivised by the QSR (without any correction for risk preference).

[22] The issue is that such stakes introduce variation in state-contingent background wealth, confounding the identification of utility and subjective beliefs. In a fully-experimental setup, the experimenter at least has knowledge of these wealth effects – since they were induced – and so the situation may potentially be salvaged through an appropriate estimation strategy. However we are not aware of any research that pursues this approach.

References

Andersen, S., Fountain, J., Harrison, G. W., Rutström, E. E., 2014. Estimating subjective probabilities. Journal of Risk and Uncertainty 48 (3), 207–229.

Andersen, S., Harrison, G. W., Lau, M. I., Rutström, E. E., 2008. Eliciting risk and time preferences. Econometrica 76 (3), 583–618.

Andreoni, J., Sprenger, C., 2012. Risk preferences are not time preferences. American Economic Review 102 (7), 3357–3376.

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

Benoît, J.-P., Dubra, J., 2011. Apparent overconfidence. Econometrica 79 (5), 1591–1625.

Benoît, J.-P., Dubra, J., Moore, D. A., 2015. Does the better-than-average effect show that people are overconfident? Two experiments. Journal of the European Economic Association 13 (2), 293–329.

Berg, J. E., Daley, L. A., Dickhaut, J. W., O'Brien, J. R., 1986. Controlling preferences for lotteries on units of experimental exchange. Quarterly Journal of Economics 101 (2), 281–306.

Bruhin, A., Santos-Pinto, L., Staubli, D., 2016. How do beliefs about skill affect risky decisions? Working paper, University of Lausanne.

Burks, S. V., Carpenter, J. P., Goette, L., Rustichini, A., 2013. Overconfidence and social signalling. Review of Economic Studies 80 (3), 949–983.

Camerer, C., Lovallo, D., 1999. Overconfidence and excess entry: An experimental approach. American Economic Review 89 (1), 306–318.

Cheung, S. L., 2015. Comment on "Risk preferences are not time preferences": On the elicitation of time preference under conditions of risk. American Economic Review 105 (7), 2242–2260.

Clark, J., Friesen, L., 2009. Overconfidence in forecasts of own performance: An experimental study. Economic Journal 119 (534), 229–251.

Daniel, K., Hirshleifer, D., 2015. Overconfident investors, predictable returns, and excessive trading. Journal of Economic Perspectives 29 (4), 61–88.

De Bondt, W. F. M., Thaler, R. H., 1995. Financial decision-making in markets and firms: A behavioral perspective. In: Jarrow, R., Maksimovic, V., Ziemba, W. T. (Eds.), Finance: Handbooks in Operations Research and Management Science. Vol. 9. pp. 385–410.

Eil, D., Rao, J. M., 2011. The good news-bad news effect: Asymmetric processing of objective information about yourself. American Economic Journal: Microeconomics 3 (2), 114–38.

Fehr-Duda, H., Epper, T., 2012. Probability and risk: Foundations and economic implications of probability-dependent risk preferences. Annual Review of Economics 4 (1), 567–593.

Fischbacher, U., 2007. z-Tree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10 (2), 171–178.

Gilboa, I., Schmeidler, D., 1989. Maxmin expected utility with non-unique prior. Journal of Mathematical Economics 18 (2), 141–153.

Gneezy, U., Potters, J., 1997. An experiment on risk taking and evaluation periods. Quarterly Journal of Economics 112 (2), 631–645.

Goldstein, W. M., Einhorn, H. J., 1987. Expression theory and the preference reversal phenomena. Psychological Review 94 (2), 236–254.

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

Goodie, A. S., 2003. The effects of control on betting: Paradoxical betting on items of high confidence with low value. Journal of Experimental Psychology: Learning, Memory, and Cognition 29 (4), 598–610.

Greiner, B., 2015. Subject pool recruitment procedures: Organizing experiments with ORSEE. Journal of the Economic Science Association 1 (1), 114–125.

Grubb, M. D., 2015. Overconfident consumers in the marketplace. Journal of Economic Perspectives 29 (4), 9–36.

Harrison, G. W., Lau, M. I., Rutström, E. E., 2013. Identifying time preferences with experiments: Comment. Working paper 2013-09, Center for the Economic Analysis of Risk, Georgia State University.

Harrison, G. W., Martínez-Correa, J., Swarthout, J. T., 2014. Eliciting subjective probabilities with binary lotteries. Journal of Economic Behavior and Organization 101, 128–140.

Heath, C., Tversky, A., 1991. Preference and belief: Ambiguity and competence in choice under uncertainty. Journal of Risk and Uncertainty 4 (1), 5–28.

Heger, S. A., Papageorge, N. W., 2016. We should totally open a restaurant: How optimism and overconfidence affect beliefs. Working paper, University of Sydney.

Hoelzl, E., Rustichini, A., 2005. Overconfident: Do you put your money on it? Economic Journal 115 (503), 305–318.

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

Karni, E., 2009. A mechanism for eliciting probabilities. Econometrica 77 (2), 603–606.

Karni, E., Safra, Z., 1995. The impossibility of experimental elicitation of subjective probabilities. Theory and Decision 38 (3), 313–320.

Kaur, S., Kremer, M., Mullainathan, S., 2015. Self-control at work. Journal of Political Economy 123 (6), 1227–1277.

Kruger, J., 1999. Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. Journal of Personality and Social Psychology 77 (2), 221–232.

Kruger, J., Windschitl, P. D., Burrus, J., Fessel, F., Chambers, J. R., 2008. The rational side of egocentrism in social comparisons. Journal of Experimental Social Psychology 44 (2), 220–232.

Malmendier, U., Tate, G., 2015. Behavioral CEOs: The role of managerial overconfidence. Journal of Economic Perspectives 29 (4), 37–60.

McFadden, D., 1974. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York, pp. 105–142.

Merkle, C., Weber, M., 2011. True overconfidence: The inability of rational information processing to account for apparent overconfidence. Organizational Behavior and Human Decision Processes 116 (2), 262–271.

Mobius, M. M., Niederle, M., Niehaus, P., Rosenblat, T. S., 2011. Managing self-confidence: Theory and experimental evidence. Working paper 17014, National Bureau of Economic Research.

Moore, D. A., Cain, D. M., 2007. Overconfidence and underconfidence: When and why people underestimate (and overestimate) the competition. Organizational Behavior and Human Decision Processes 103 (2), 197–213.

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

Murad, Z., Sefton, M., Starmer, C., 2016. How do risk attitudes affect measured confidence? Journal of Risk and Uncertainty 52 (1), 21–46.

Niederle, M., Vesterlund, L., 2007. Do women shy away from competition? Do men compete too much? Quarterly Journal of Economics 122 (3), 1067–1101.

Offerman, T., Sonnemans, J., Van De Kuilen, G., Wakker, P. P., 2009. A truth serum for non-Bayesians: Correcting proper scoring rules for risk attitudes. Review of Economic Studies 76 (4), 1461–1489.

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

Price, P. C., 1998. Effects of a relative-frequency elicitation question on likelihood judgment accuracy: The case of external correspondence. Organizational Behavior and Human Decision Processes 76 (3), 277–297.

Quiggin, J., 1982. A theory of anticipated utility. Journal of Economic Behavior and Organization 3 (4), 323–343.

Schlag, K. H., Tremewan, J., van der Weele, J. J., 2015. A penny for your thoughts: A survey of methods for eliciting beliefs. Experimental Economics 18 (3), 457–490.

Svenson, O., 1981. Are we all less risky and more skillful than our fellow drivers? Acta Psychologica 47 (2), 143–148.

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

Weinstein, N. D., 1980. Unrealistic optimism about future life events. Journal of Personality and Social Psychology 39 (5), 806–820.

Winkler, R. L., Murphy, A. H., 1970. Nonlinear utility and the probability score. Journal of Applied Meteorology 9 (1), 143–148.

Appendices: For online publication only

A Additional tables and figures

Table A.1: Spearman rank correlation matrix of summary variables (full sample)

	Age	Payment	Skill (Trial)	Skill (Main)	Allocation: Risk	Allocation: Placement
Age	1					
Payment	0.028	1				
	(0.764)					
Skill (Trial)	–0.110	0.544 ***	1			
	(0.230)	(0.000)				
Skill (Main)	–0.015	0.663 ***	0.784 ***	1		
	(0.867)	(0.000)	(0.000)			
Allocation: Risk	–0.015	0.007	–0.036	–0.058	1	
	(0.873)	(0.938)	(0.699)	(0.526)		
Allocation: Placement	0.027	0.267 ***	0.577 ***	0.358 ***	0.188 **	1
	(0.771)	(0.003)	(0.000)	(0.000)	(0.040)	

Notes: p-values in parentheses. ** $p < 0.05$; *** $p < 0.01$.

Table A.2: Summary statistics (by treatment)

	Mean	SD	Median	10th pctl	90th pctl
T1: Low information					
Female	0.375				
Age	22.325	3.100	21.500	19.000	25.000
Payment	31.300	20.322	25.000	12.000	59.000
Skill task score (Trial round)	12.650	6.045	11.000	7.000	19.500
Skill task score (Main round)	13.150	6.066	11.500	9.000	21.000
Mean allocation: Risk	0.423	0.199	0.400	0.167	0.663
Mean allocation: Placement	0.525	0.320	0.536	0.059	0.999
T2: Medium information					
Female	0.475				
Age	24.250	5.148	23.000	20.000	32.500
Payment	29.700	12.396	27.500	17.500	46.500
Skill task score (Trial round)	11.350	3.520	11.000	6.500	16.500
Skill task score (Main round)	12.975	3.833	12.500	8.500	16.500
Mean allocation: Risk	0.389	0.184	0.388	0.136	0.646
Mean allocation: Placement	0.420	0.329	0.453	0.000	0.844
T3: High information					
Female	0.500				
Age	21.125	2.483	21.000	18.000	24.000
Payment	32.150	17.634	27.000	17.000	49.500
Skill task score (Trial round)	10.950	4.744	11.000	6.500	14.000
Skill task score (Main round)	13.150	5.284	12.000	8.500	17.000
Mean allocation: Risk	0.379	0.164	0.384	0.149	0.602
Mean allocation: Placement	0.420	0.371	0.357	0.000	1.000

Table A.3: Tests of treatment differences in summary variables (two-sided p-values)

	Joint	T1 vs T2	T1 vs T3	T2 vs T3
Female	0.508	0.498	0.367	1.000
Age	0.004 ***	0.081 *	0.066 *	0.002 ***
Payment	0.742	0.567	0.467	0.851
Skill task score (Trial round)	0.444	0.636	0.233	0.374
Skill task score (Main round)	0.519	0.286	0.706	0.408
Mean allocation: Risk	0.616	0.498	0.336	0.802
Mean allocation: Placement	0.211	0.138	0.121	0.828

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Tests of differences in gender are Fisher exact tests. For remaining variables, tests in the three treatments jointly are Kruskal-Wallis tests, while pairwise comparisons are in Wilcoxon rank-sum tests.

Table A.4: Individual MNL parameter estimates (T1: Low information)

	Mean	SD	Median	10th pctl	90th pctl
Expected utility					
θ: utility curvature	1.186	3.133	0.619	0.078	2.178
q: subjective belief	0.592	0.359	0.644	0.008	1.000
λ: scale	0.052	0.067	0.032	0.009	0.094
Rank-dependent utility (Goldstein-Einhorn)					
θ: utility curvature	2.086	6.728	0.266	–0.565	3.039
γ: weighting, curvature	1.979	2.723	0.559	0.003	6.085
δ: weighting, elevation	2.089	3.325	0.590	0.252	9.751
q: subjective belief	0.625	0.363	0.722	0.047	1.000
λ: scale	0.125	0.602	0.006	0.000	0.074
Rank-dependent utility (Prelec)					
θ: utility curvature	0.897	2.818	0.081	–0.618	3.040
α: weighting, curvature	1.840	3.009	0.297	0.000	7.004
β: weighting, elevation	1.728	2.065	1.099	0.462	4.156
q: subjective belief	0.655	0.343	0.786	0.131	1.000
λ: scale	0.012	0.023	0.002	0.000	0.039
Direct statement					
q: subjective belief	0.636	0.252	0.675	0.325	0.995

Table A.5: Individual MNL parameter estimates (T2: Medium information)

	Mean	SD	Median	10th pctl	90th pctl
Expected utility					
θ: utility curvature	1.414	4.729	0.687	–0.245	2.286
q: subjective belief	0.482	0.386	0.450	0.000	1.000
λ: scale	0.209	0.661	0.046	0.008	0.233
Rank-dependent utility (Goldstein-Einhorn)					
θ: utility curvature	1.444	4.220	0.177	0.000	2.759
γ: weighting, curvature	0.972	1.863	0.302	0.000	3.214
δ: weighting, elevation	1.689	2.857	0.504	0.371	6.284
q: subjective belief	0.487	0.368	0.484	0.000	1.000
λ: scale	0.010	0.020	0.001	0.000	0.029
Rank-dependent utility (Prelec)					
θ: utility curvature	1.717	4.115	0.317	0.000	5.640
α: weighting, curvature	1.604	2.757	0.474	0.000	7.054
β: weighting, elevation	1.510	1.495	1.099	0.211	3.726
q: subjective belief	0.475	0.376	0.501	0.000	1.000
λ: scale	0.103	0.541	0.001	0.000	0.050
Direct statement					
q: subjective belief	0.527	0.277	0.500	0.150	0.925

Table A.6: Individual MNL parameter estimates (T3: High information)

	Mean	SD	Median	10th pctl	90th pctl
Expected utility					
θ: utility curvature	1.027	1.229	0.762	0.258	2.248
q: subjective belief	0.511	0.401	0.481	0.000	1.000
λ: scale	0.061	0.139	0.028	0.005	0.123
Rank-dependent utility (Goldstein-Einhorn)					
θ: utility curvature	0.820	1.590	0.104	0.000	2.597
γ: weighting, curvature	0.682	1.595	0.091	0.000	1.338
δ: weighting, elevation	1.137	1.655	0.516	0.495	2.122
q: subjective belief	0.520	0.396	0.533	0.000	1.000
λ: scale	0.004	0.006	0.002	0.000	0.012
Rank-dependent utility (Prelec)					
θ: utility curvature	2.525	5.580	0.262	0.000	8.620
α: weighting, curvature	1.488	2.633	0.354	0.000	5.526
β: weighting, elevation	1.189	1.279	1.089	0.334	1.830
q: subjective belief	0.470	0.412	0.415	0.000	1.000
λ: scale	0.020	0.103	0.001	0.000	0.011
Direct statement					
q: subjective belief	0.526	0.307	0.500	0.100	1.000

Table A.7: Significance of individual risk preference parameter estimates

	MNL-EU	MNL-GE	MNL-P2
Concave	104	92	95
Convex	3	1	0
Inverse-S		17	14
S-shaped		19	24
Optimistic		6	13
Pessimistic		26	39
Weighting, joint		39	60

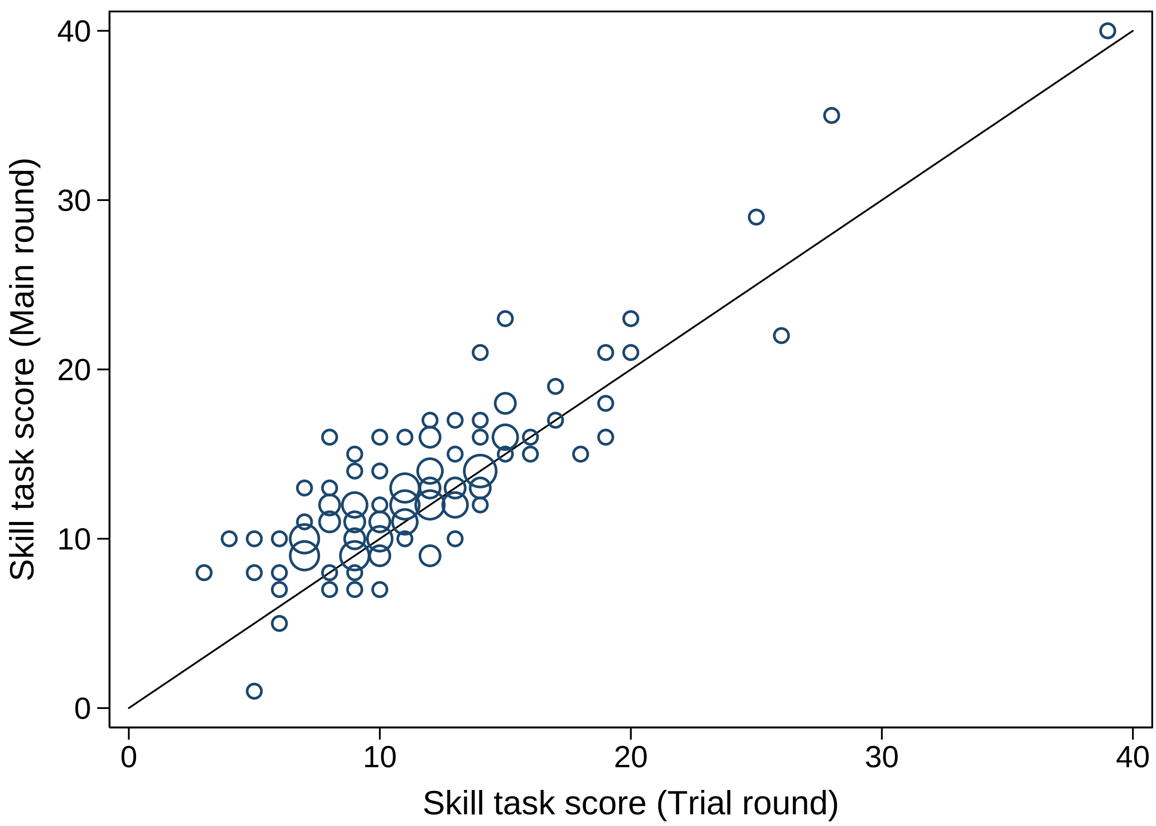
Notes: This table summarises the number of subjects whose individual utility and probability weighting parameter estimates differ significantly from linearity at the 5% level. Concave utility corresponds to $\theta > 0$. Inverse-S probability weighting corresponds to $\gamma < 1$ in the GE specification, or $\alpha < 1$ in the P2 specification. Optimism corresponds to $\delta > 1$ in the GE specification, or $\beta < 1$ in the P2 specification. The final row shows subjects for whom the two probability weighting parameters are jointly significant.

Table A.8: Comparison of estimated beliefs to direct reports (p-values)

	Full sample	T1: Low info	T2 Med info	T3 High info
Levels (Wilcoxon signed-rank tests)				
MNL-EU	0.249	0.368	0.563	0.667
MNL-GE	0.624	0.979	0.554	0.697
MNL-P2	0.305	0.648	0.397	0.119
Categorical (Fisher exact tests)				
MNL-EU	0.000 ***	0.213	0.040 **	0.037 **
MNL-GE	0.001 ***	0.071 *	0.168	0.084 *
MNL-P2	0.000 ***	0.078 *	0.090 *	0.037 **

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. The tests in the top panel compare estimated to directly-reported beliefs in Wilcoxon signed-rank tests, confirming no significant difference in levels. In the bottom panel we collapse the beliefs into three categories: $q \in [0, 0.1]$, $(0.1, 0.9)$, $[0.9, 1]$ and test whether the distribution of beliefs over these categories differs between estimated and directly-reported beliefs using Fisher exact tests.

Figure A.1: Individual performance in the skill task, trial versus main round



B Alternative estimation procedures

The alternative estimation procedures used as robustness checks are nonlinear least squares (NLS) and a Tobit model estimated by maximum likelihood. Both procedures estimate the individual risk preference parameters (according to specification) and subjective belief q_i using the solution function to a subject's investment allocation problem. We derive this solution function below for the more complex placement decisions; analogous arguments apply for the more straightforward risk decisions.

Subject i's maximisation problem in the jth placement decision is:

$$\max_s RDU^P\left(s; p_j, q_i\right) \quad s.t. \quad s \in [0,1], \tag{B.1}$$

where RDU^P is defined by equation (3) and utility by equation (1) in the text. The first-order condition for this problem is:

$$\frac{2w\left(p_j q_i\right)}{1-w\left(1-\left(1-p_j\right)\left(1-q_i\right)\right)} = \left(\frac{1+2s}{1-s}\right)^{\theta_i}. \tag{B.2}$$

The first-order condition in this form implies that there is, at most, one value of $s \in \left[-\frac{1}{2}, 1\right)$ that is at a stationary point as the RHS is strictly positive and increasing in s over this interval. Moreover, there must be such a value as both sides of the equation are continuous functions over the same range $[0, \infty)$.

The stationary point of $s \in \left[-\frac{1}{2}, 1\right)$ is always a maximum for $\theta_i > 0$ since:

$$\frac{\partial^2 RDU^P}{\partial s^2} = -\theta_i \left[\frac{4w\left(p_j q_i\right)}{(1+2s)^{\theta_i+1}} + \frac{1-w\left(1-\left(1-p_j\right)\left(1-q_i\right)\right)}{(1-s)^{\theta_i+1}}\right] < 0. \tag{B.3}$$

It is clear from equation (3) in the text that RDU is continuous over the same interval, $s \in \left[-\frac{1}{2}, 1\right)$. Since there is only one stationary point over this interval, which is a maximum, RDU is strictly decreasing as s moves further away from the stationary point. Therefore, i's optimal allocation in decision j, s^*_{ij}, is at the stationary point if this satisfies $s \in [0,1]$ and 0 otherwise:

$$\left(s^*_{ij}\middle| j \in \{10, \ldots, 18\}, \theta_i > 0\right) = \max\left\{0, \frac{\left(\frac{2w\left(p_j q_i\right)}{1-w\left(1-\left(1-p_j\right)\left(1-q_i\right)\right)}\right)^{\frac{1}{\theta_i}} - 1}{\left(\frac{2w\left(p_j q_i\right)}{1-w\left(1-\left(1-p_j\right)\left(1-q_i\right)\right)}\right)^{\frac{1}{\theta_i}} + 2}\right\}. \tag{B.4}$$

The value of s^* derived above will be a minimum for $\theta_i < 0$ and an inflection point for $\theta_i = 0$. Any individuals with these preferences will always maximise utility at a boundary point. The condition for selecting the upper boundary ($s^*_{ij} = 1$) is:

$$w\left(p_j q_i\right)\left(3^{1-\theta_i} - 1\right) + w\left(1 - \left(1 - p_j\right)\left(1 - q_i\right)\right) > 1. \tag{B.5}$$

Similar arguments apply for s^*_{ij} in the risk decisions. The solution function for the case of concave utility is given by:

$$\left(s^*_{ij}\middle| j \in \{1, \ldots, 9\}, \theta_i > 0\right) = \max\left\{0, \frac{\left(\frac{2w\left(p_j\right)}{1 - w\left(p_j\right)}\right)^{\frac{1}{\theta_i}} - 1}{\left(\frac{2w\left(p_j\right)}{1 - w\left(p_j\right)}\right)^{\frac{1}{\theta_i}} + 2}\right\}, \tag{B.6}$$

while for the case of $\theta_i \leq 0$ the upper boundary ($s^*_{ij} = 1$) is selected if:

$$w\left(p_j\right)\left(3^{1-\theta_i} - 1\right) > 1. \tag{B.7}$$

B.1 Least-squares estimation

The least-squares estimator is defined as:

$$\left(\hat{\Theta}_i, \hat{q}_i\right) = \arg\min_{(\Theta_i, q_i)} \sum_{j=1}^{18} \left(s_{ij} - s^*_{ij}\right)^2, \tag{B.8}$$

where Θ_i is the vector of individual risk-preference parameters, according to specification: $\Theta_i^{EU} = \{\theta_i\}$, $\Theta_i^{GE} = \{\theta_i, \gamma_i, \delta_i\}$, or $\Theta_i^{P2} = \{\theta_i, \alpha_i, \beta_i\}$.

B.2 Tobit estimation

In the Tobit model, the optimal allocations s^T_{ij} are modified slightly by the addition of a decision error term, $\varepsilon \sim N\left(0, \sigma^2\right)$. For the case of concave utility $\theta_i > 0$:

$$\left(s^T_{ij}\middle| \theta_i > 0\right) = \max\left\{0, \tilde{s}_{ij} + \varepsilon_{ij}\right\}, \tag{B.9}$$

$$\tilde{s}_{ij} = \begin{cases} \dfrac{\left(\dfrac{2w\left(p_j\right)}{1-w\left(p_j\right)}\right)^{\frac{1}{\theta_i}} - 1}{\left(\dfrac{2w\left(p_j\right)}{1-w\left(p_j\right)}\right)^{\frac{1}{\theta_i}} + 2} & \text{if } j \in \{1, \ldots, 9\}, \\ \dfrac{\left(\dfrac{2w\left(p_j q_i\right)}{1-w\left(1-\left(1-p_j\right)\left(1-q_i\right)\right)}\right)^{\frac{1}{\theta_i}} - 1}{\left(\dfrac{2w\left(p_j q_i\right)}{1-w\left(1-\left(1-p_j\right)\left(1-q_i\right)\right)}\right)^{\frac{1}{\theta_i}} + 2} & \text{if } j \in \{10, \ldots, 18\}. \end{cases} \tag{B.10}$$

For the case of $\theta_i \leq 0$:

$$\left(s_{ij}^T \middle| \, \theta_i \leq 0\right) = \begin{cases} 1 & \text{if} \quad w\left(p_j\right)\left(3^{1-\theta_i} - 1\right) > 1 \text{ and } j \in \{1, \ldots, 9\}, \\ & \text{or} \quad w\left(p_j q_i\right)\left(3^{1-\theta_i} - 1\right) + w\left(1-\left(1-p_j\right)\left(1-q_i\right)\right) > 1 \\ & \qquad \text{and } j \in \{10, \ldots, 18\}, \\ 0 & \text{otherwise.} \end{cases} \tag{B.11}$$

The probability function of an observed allocation is defined as:

$$\Pr\left(s_{ij}\right) = \begin{cases} \Phi\left(-\tilde{s}_{ij}\right) & \text{if } \theta_i > 0, \;\; s_{ij} = 0, \\ \phi\left(s_{ij} - \tilde{s}_{ij}\right) & \text{if } \theta_i > 0, \;\; 0 < s_{ij} < 1, \\ 1 - \Phi\left(1 - \tilde{s}_{ij}\right) & \text{if } \theta_i > 0, \;\; s_{ij} = 1, \\ 1 & \text{if } \theta_i \leq 0, \;\; s_{ij} = s_{ij}^T, \\ 0 & \text{if } \theta_i \leq 0, \;\; s_{ij} \neq s_{ij}^T, \end{cases} \tag{B.12}$$

where ϕ is the PDF and Φ the CDF of the error term, ε.

The log-likelihood function is then constructed by summing the log probabilities across j:

$$LL_i^T = \sum_{j=1}^{18} \ln \Pr\left(s_{ij}\right), \tag{B.13}$$

with maximum likelihood parameters:

$$\left(\hat{\Theta}_i, \hat{q}_i, \hat{\sigma}_i\right) = \arg \max_{\left(\Theta_i, q_i, \sigma_i\right)} LL_i^T. \tag{B.14}$$

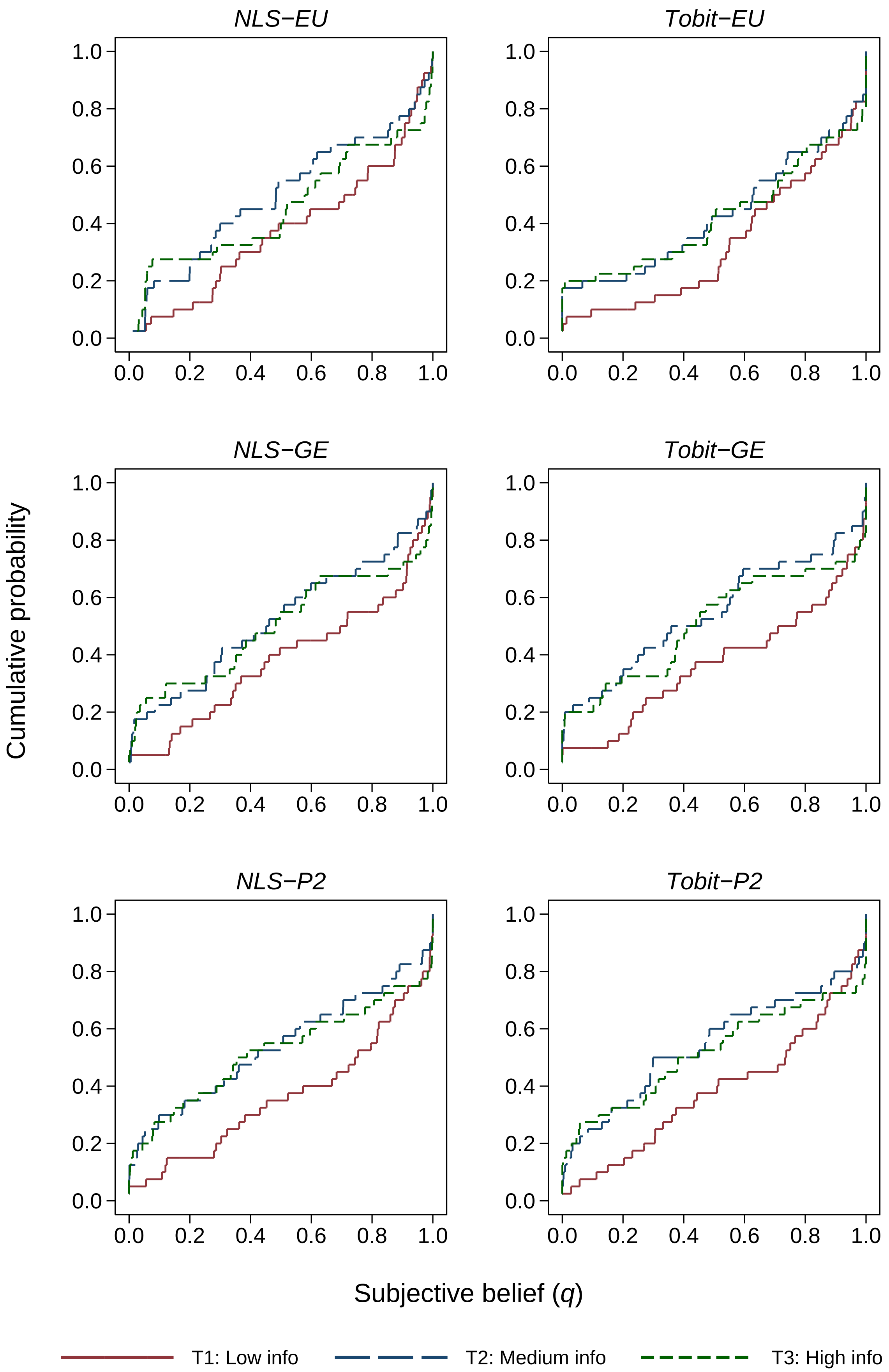
Table B.1: Individual NLS parameter estimates (full sample)

	Mean	SD	Median	10th pctl	90th pctl
	Expected utility				
θ: utility curvature	1.410	3.056	0.892	0.296	2.321
q: subjective belief	0.560	0.345	0.586	0.054	0.989
	Rank-dependent utility (Goldstein-Einhorn)				
θ: utility curvature	2.362	3.887	1.325	0.001	5.169
γ: weighting, curvature	1.868	2.315	0.970	0.001	5.765
δ: weighting, elevation	4.316	4.080	2.020	0.500	10.000
q: subjective belief	0.531	0.360	0.500	0.016	0.992
	Rank-dependent utility (Prelec)				
θ: utility curvature	2.475	3.672	0.747	0.000	7.229
α: weighting, curvature	2.067	2.728	0.797	0.001	6.696
β: weighting, elevation	1.222	1.653	1.049	0.004	2.131
q: subjective belief	0.525	0.372	0.535	0.005	0.997

Table B.2: Individual Tobit parameter estimates (full sample)

	Mean	SD	Median	10th pctl	90th pctl
	Expected utility				
θ: utility curvature	2.956	5.749	1.110	0.473	4.960
q: subjective belief	0.606	0.350	0.682	0.000	1.000
σ: error	1.168	6.694	0.218	0.098	0.599
	Rank-dependent utility (Goldstein-Einhorn)				
θ: utility curvature	2.167	3.566	1.188	0.000	4.888
γ: weighting, curvature	1.754	2.226	0.909	0.000	5.374
δ: weighting, elevation	3.683	3.461	2.661	0.500	9.997
q: subjective belief	0.525	0.370	0.519	0.001	1.000
σ: error	0.400	2.419	0.101	0.044	0.223
	Rank-dependent utility (Prelec)				
θ: utility curvature	2.518	3.776	0.797	0.000	7.366
α: weighting, curvature	2.105	2.692	0.932	0.000	6.123
β: weighting, elevation	1.030	1.410	0.999	0.001	1.336
q: subjective belief	0.519	0.370	0.497	0.011	1.000
σ: error	0.342	2.081	0.094	0.038	0.246

Figure B.1: CDFs of NLS and Tobit estimates of individual beliefs (q), by model and treatment



C Instructions for the low information treatment

Thank you for participating in this experiment. The experiment today will be done through the computers on your desks. Please note that once the experiment begins you should not speak to the other participants. If you have any questions throughout the session, raise your hand and wait for the supervisor. Please also switch your mobile phone off and place it in your bag. All of the items you require for this experiment are provided to you on the desks. *These rules are important for the integrity of the experiment and you may be asked to leave without payment if you do not follow them.*

Today we will be asking you to complete an addition task for payment. You will also have the opportunity to allocate your earnings from this task between different accounts. At the end of the session, these accounts will pay out different amounts, which may depend on the performance of the group in the addition task as well as chance. Full information on these accounts will be provided later.

The order of the experiment will be:

1. Addition task first round
2. First allocation decisions
3. Second allocation decisions with feedback from the first round of the addition task
4. Addition task second round

Each of these stages will be explained fully to you in separate instructions as you progress through the experiment.

Payment

Your final payment will depend on the decisions you make in at least one of the above stages. You will only know which stages are relevant to your payment after making all of your decisions, so think through each section carefully. At the end of the experiment, the computer will randomly decide your payment method out of three equally likely possibilities:

Payment Method 1. 1/3 chance – Flat payment rate for first round performance: If you are selected to be paid in this way, you will receive $2.50 for every correctly answered addition question in stage 1, the addition task first round.

Payment Method 2. 1/3 chance – Earnings from second round performance and one randomly chosen allocation decision from the first allocation stage: If you are paid this way, your base earnings will be determined by your performance in the second round of the addition task (stage 4). You would receive base earnings of $1.50 for every correctly answered question in stage 4. The computer would also randomly select one of the allocation decisions you made in the first set of allocation decisions (stage 2) to apply to your base earnings. You will be given more information on the first allocation stage when you reach stage 2.

Payment Method 3. 1/3 chance – Earnings from second round performance and one randomly chosen allocation decision from the second allocation stage: If you are paid this way, your base earnings will be determined by your performance in the second round of the addition task (stage 4). You would receive base earnings of $1.50 for every correctly answered question in stage 4. The computer would also randomly select one of the allocation decisions you made in the second set of allocation decisions (stage 3) to apply to your base earnings. You will be given more information on the second allocation stage when you reach stage 3.

You will also receive $5.00 in addition to your earnings in the experiment for completing the session. The total of your experiment earnings and $5.00 participation fee will be given to you in cash at the end of the session.

Stage 1 – Addition task first round

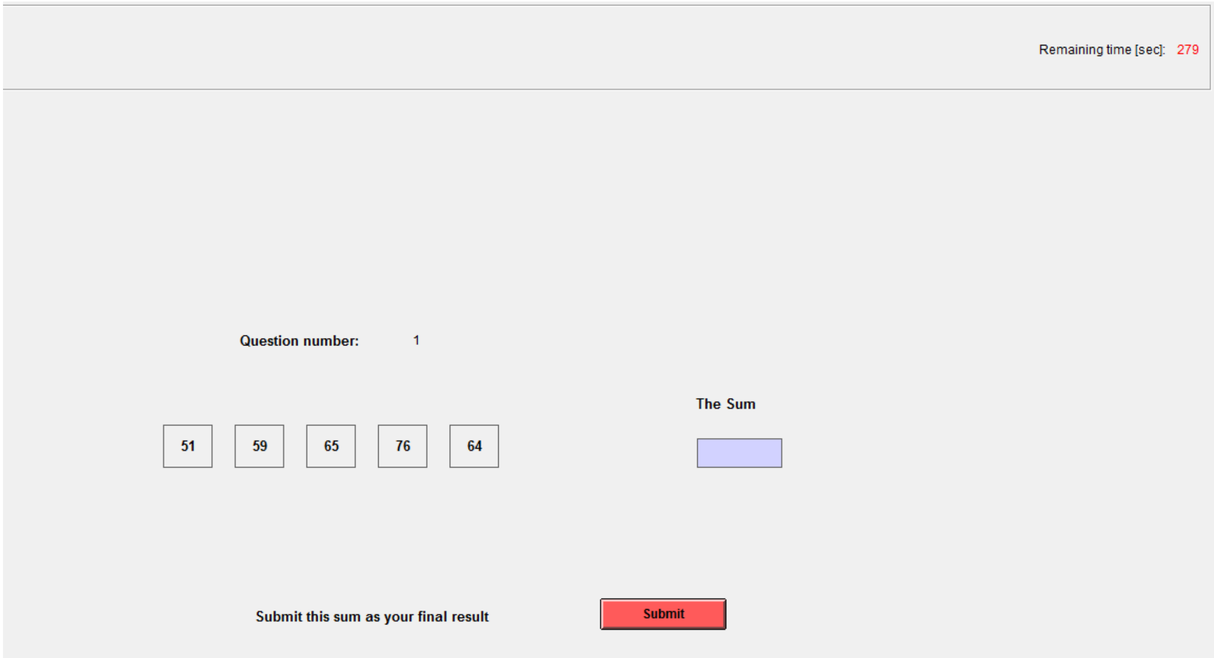
You will soon complete a first round of the addition task. This is so you and the other participants have an idea of your performance before the rest of the experiment. If you are selected to be paid through payment method 2 or 3, this first round will be practice for the second round, where you will earn \$1.50 for each correct answer. However, it is in your interest to try your best in the first round as there is a 1/3 chance that the computer will randomly select you at the end to be paid according payment method 1. If you are selected for this payment method, you will receive \$2.50 for every correctly answered addition question in the first round of the addition task.

The task that you will complete is a simple addition task. You will be presented with sets of five 2-digit numbers and required to add as many of them as you can in 5 minutes. You are free to use working paper for these questions but *you are not permitted to use a calculator*.

Example question

$76 + 78 + 42 + 44 + 43 = ?$

Answer = 283



The five double-digit numbers for you to add will be shown to you in individual boxes in the centre of the screen. To complete each addition question, enter your answer in the input box labelled "The Sum" and click the submit button. After submitting your answer, you will immediately move to the next screen with another set of five numbers to add. There is no limit to the number of questions you can answer. The task will only end once five minutes have elapsed. You will be told how many questions you answered correctly in stage 3. There is no penalty for entering an incorrect answer.

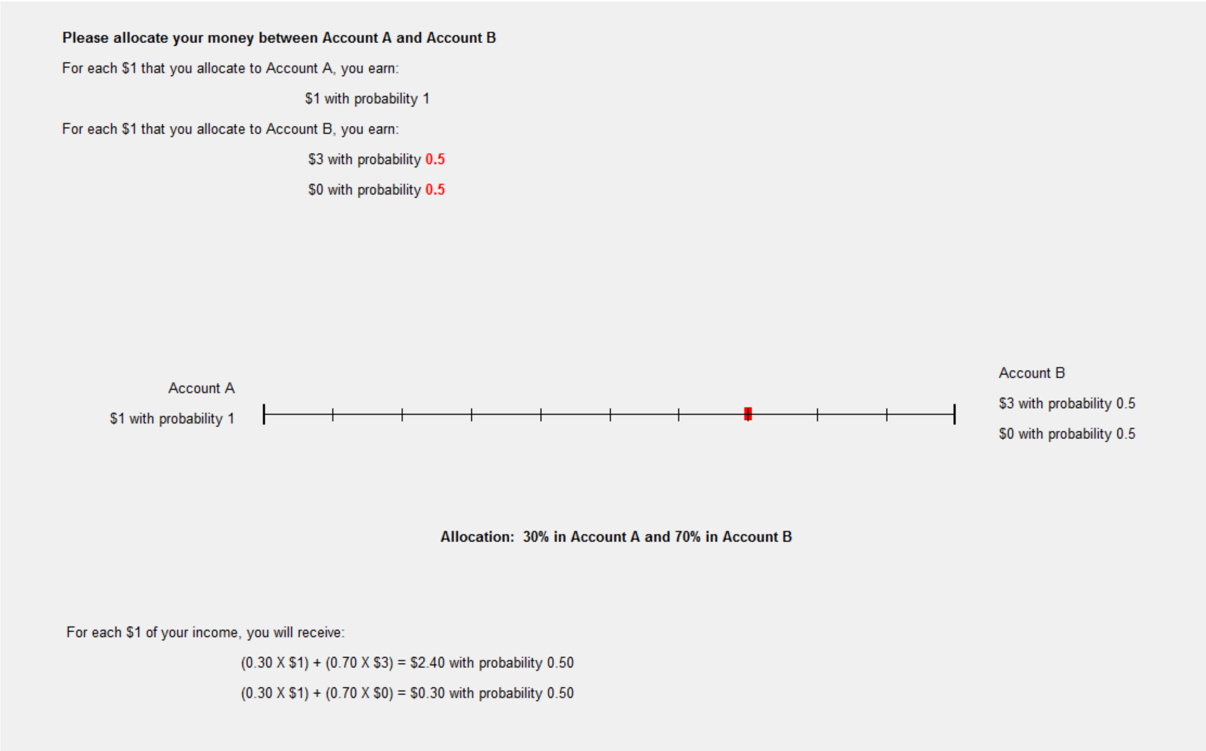
Stage 2 – First allocation decisions

In this stage, you will be asked to allocate the money that you will earn in the second round of the addition task between two accounts, Account A and Account B. Account A will pay you \$1 for every \$1 that you allocate to it with probability 1. Account B is a risky account. It will have a chance of paying you \$3 for every \$1 that you allocate to it but also a chance of paying \$0.

You will be told the probability of Account B paying \$3 per \$1 on the computer screen before making your decision. There will be a series of these decisions to make with different probabilities of paying \$3 and you can allocate between the accounts differently in each case. You should consider each of these decisions carefully as one will be selected at random to determine your final payment if you are paid through payment method 2.

You will not know exactly how much money you will have available when you make these choices as this would depend on your performance in the second round of the task. For this reason, we will ask you to specify percentages to allocate rather than dollar amounts.

Example:



For each $1 that you allocate to Account A, it will pay you $1 with probability 1.

In this example, for each $1 that you allocate to Account B, it will pay you:

$3 with probability 0.5
$0 with probability 0.5

Allocate your future earnings between Account A and Account B by placing or dragging the red slider on the bar in the centre of the screen with the mouse. The slider will only appear once you have clicked to place it on the bar. Moving the slider closer to the right allocates a greater percentage of your money to Account B, up to 100% at the right-hand end of the bar. You can proceed to the next decision or go back with the buttons in the bottom right corner. Once you complete all of the decisions, a submit button will appear.

Suppose that you decide to allocate 30% of your earnings to Account A and the remaining 70% to Account B. After making this decision, suppose that you earn $20 in the second round of the addition task. Your payment will therefore be:

$(0.3 \times \$20 \times 1) + (0.7 \times \$20 \times 3) = \$48$ if Account B pays $3 per $1 or;
$(0.3 \times \$20 \times 1) + (0.7 \times \$20 \times 0) = \$6$ if Account B pays $0 per $1

Note that this does not include your $5 participation payment, which is added at the end.

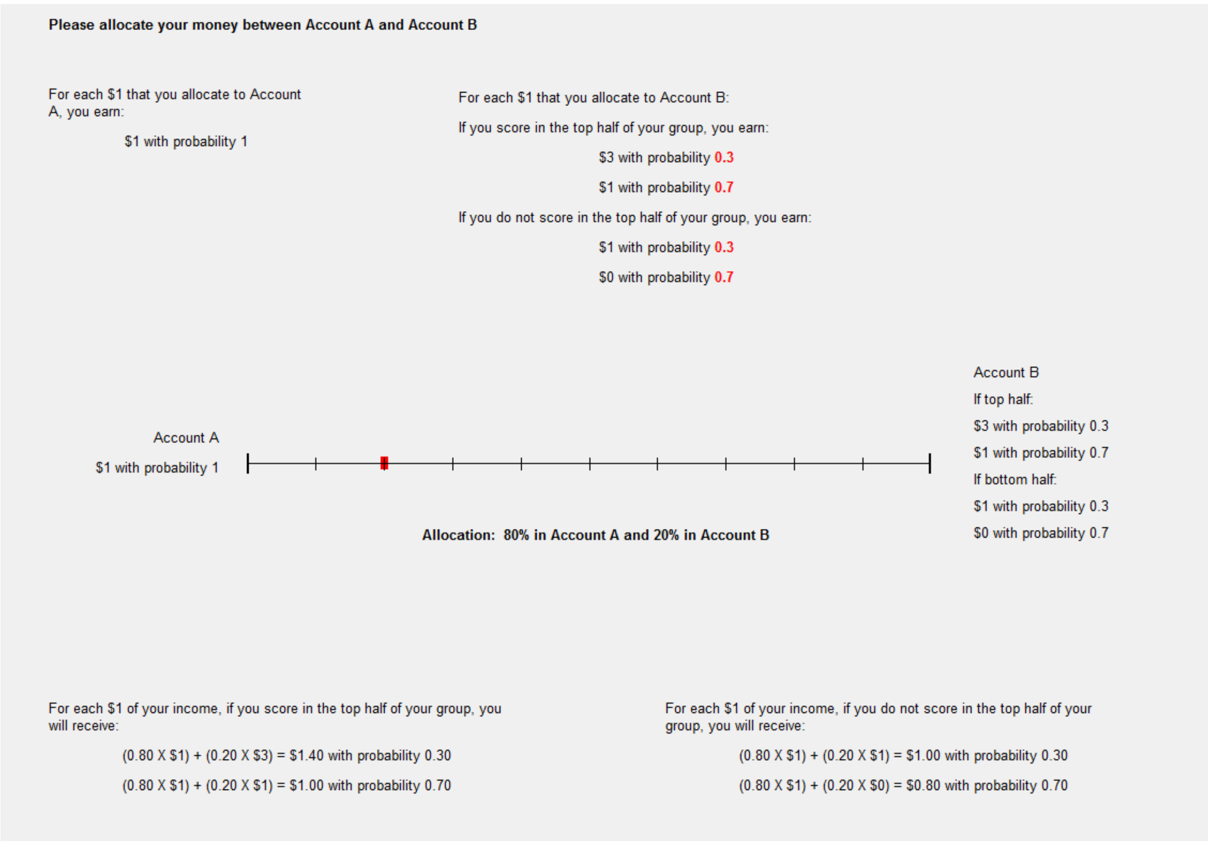
Stage 3 – Second allocation decisions with feedback from first addition task

In this stage, you will go through a second series of allocation decisions. These decisions are similar to before but with one important difference in Account B. The payout from Account B will now depend on your ranking *in the second round of the addition task* within a sub-group of 4 people, as well as chance. You have each been randomly assigned to one of these sub-groups of 4 people out of all of the participants here today.

Recall from stage 2 that Account B paid out either \$3 or \$0 for every \$1 that you allocated to it. Now, if you score in the top half of your group (top two of the group of four), Account B will pay out either \$3 or \$1 for every \$1 you allocate to it. If you score in the bottom half, Account B will pay out either \$1 or \$0 for every \$1 you allocate to it. If there is a tie within your group, the computer will randomly break the tie, e.g. if two group members tie for second place, the computer will randomly choose one of them to place second and the other to place third. As in the first allocation decisions, the computer program will inform you of the probabilities of each amount that Account B may pay before you make your decision. Account A will still pay \$1 for every \$1 that you allocate to it with probability 1.

Once again, you will make a series of these decisions for different probabilities of each amount that Account B may pay. Each of these decisions has an equal chance under payment method 3 of being chosen to determine your final payment.

To help you consider your chances of scoring in the top half of your group of 4 in the upcoming second round of the addition task, you will be shown a feedback screen before making the allocation decisions. You will be told: how many questions you answered correctly in the first round of the addition task and how many questions you attempted in total.



Example:

For each \$1 that you allocate to Account A, it will pay you \$1 with probability 1.

In this example, for each \$1 that you allocate to Account B, it will pay you:

If you score in the top half of your group:

\$3 with probability 0.3
\$1 with probability 0.7

If you do not score in the top half:

\$1 with probability 0.3
\$0 with probability 0.7

Suppose that you decide to allocate 80% of your earnings to Account A and the remaining 20% to Account B.

After making this decision, suppose that you earn \$15 in second round of the addition task and score in the top half of your group. Your payment will therefore be:

$(0.8 \times \$15 \times 1) + (0.2 \times \$15 \times 3) = \$21$ if Account B pays \$3 per \$1 or;
$(0.8 \times \$15 \times 1) + (0.2 \times \$15 \times 1) = \$15$ if Account B pays \$1 per \$1

If you had scored in the bottom half, your possible payment amounts would have been:

$(0.8 \times \$15 \times 1) + (0.2 \times \$15 \times 1) = \$15$ if Account B pays \$1 per \$1 or;
$(0.8 \times \$15 \times 1) + (0.2 \times \$15 \times 0) = \$12$ if Account B pays \$0 per \$1

Stage 4 – Addition task second round

You will now complete the second round of the addition task. If you are selected to be paid by payment method 2 or 3, you will earn \$1.50 for each addition question you correctly answer. These earnings will then be allocated according to a randomly chosen allocation decision. The task itself will be the same format as the first round. Only the numbers that you are required to add will be different. The same time limit of 5 minutes applies.