Does the Better-Than-Average Effect Show That People Are Overconfident?: Two Experiments.[*]

Jean-Pierre Benoît[†]
London Business School

Juan Dubra
Universidad de Montevideo

Don Moore
Haas School of Business, UC Berkeley.

Abstract

We conduct two experimental tests of the claim that people are overconfident, using new tests of overplacement that are based on a formal Bayesian model. Our two experiments, on easy quizzes, find that people overplace themselves. More precisely, we find apparently overconfident data that cannot be accounted for by a rational population of expected utility maximizers who care only about money. The finding represents new evidence of overconfidence that is robust to the Bayesian critique offered by Benoît and Dubra (2011). We discuss possible limitations of our results.

Keywords: Overconfidence; Better than Average; Experimental Economics; Irrationality; Signalling Models.

Journal of Economic Literature Classification Numbers: D11, D12, D82, D83

1 Introduction

A large body of literature across several disciplines, including psychology, finance, and economics, purports to find that people are generally overconfident.[1] While the term overconfidence has been used in a variety of ways, we focus here on the phenomenon of people overplacing themselves relative to others (Larrick, Burson, and Soll, 2007). One manifestation of this

[*] Authors appear in alphabetical order. This paper was previously circulated as 'A Proper Test of Overconfidence.' We thank Uriel Haran for help with data collection as well as the staff and facilities of the Center for Behavioral Decision Research at Carnegie Mellon University. Data and materials presented in this paper are available at learnmoore.org/mooredata/OJD/

[†] jpbenoit@london.edu

[1] Papers on overconfidence in economics include Camerer and Lovallo (1999) analyzing entry in an industry, Fang and Moscarini (2005) analyzing the effect of overconfidence on optimal wage setting, Garcia, Sangiorgi and Urosevic (2007) analyzing the efficiency consequences of overconfidence in information acquisition in

overplacement is the so-called better-than-average effect, in which a strict majority of people claim to be more adept in some domain than half the population. This effect has been noted for a wide range of easy endeavours, from driving skill to spoken expression to the ability to get along with others.[2] While this effect is well-established, Benoît and Dubra (2011–henceforth B&D) have questioned its meaning. They show that, although better-than-average data gives the appearance that (some) people are overplacing themselves, it does not, *in and of itself*, indicate true overplacement, or overconfidence, which carries the implication that people have made an error in their self-evaluations.[3] This concern applies to the majority of the prior literature on the better-than-average effect.

In this paper, we report on two experiments that use new tests of overconfidence. These experiments also find that subjects overplace themselves, but in a way that cannot easily be accounted for by error-free rational agents. In this sense, these experiments find true overconfidence. While theoretically sound, these experiments are nonetheless subject to their own caveats, to which we will return.

1.1 Background

The most common type of study in this field asks subjects, either explicitly or implicitly, how they rank relative to others. For instance, Svenson (1981), in perhaps the most cited study, asks his subjects to estimate how their driving compares to others by placing themselves into one of ten deciles, while Hoelzl and Rustichini (2005) obtain implicit rankings by asking subjects if they are willing to bet that they will score in the top half of their group on a vocabulary quiz.

Suppose that over 50% of subjects rank themselves in the top half. For the most part, experimenters have simply asserted that this apparent overplacement is evidence of true overconfidence without a careful articulation of why this is so. The following example shows that this approach can be problematic.

Take a large population of subjects who are asked to rank themselves on their ability to solve logic puzzles. Suppose that, as it happens, 55% of the population can be classified as

financial markets, Kőszegi (2006) who studies how overconfidence affects the way people choose tasks or careers, and Menkhoff et al. (2006) who analyze the effect of overconfidence on herding by fund managers. In finance, papers include Barber and Odean (2001), Bernardo and Welch (2001), Chuang and Lee (2006), Daniel, Hirshleifer and Subrahmanyam (2001), Kyle and Wang (1997), Malmendier and Tate (2005), Peng and Xiong (2006), and Wang (2001). See Benoît and Dubra (2011) for a discussion of some of the literature.

[2] Early research pointed towards a universal better-than-average effect but more recent work indicates that the effect is primarily for easy tasks and may be reversed for difficult tasks.

[3] Other papers which question the significance of the better-than-average effect include Zábojník (2004) and Brocas and Carillo (2007). Harris and Hahn (2011) argue that many of the empirical findings are statistical artefacts.

low-skilled, 25% medium-skilled, and 20% high-skilled.[4] For simplicity, suppose that subjects initially have no specific information on how adept they personally are at this particular type of logic puzzles. However, they are given two sample puzzles to solve. The puzzles are such that a low-skilled person has an (i.i.d.) 0.1 chance of solving a puzzle correctly, a medium-skilled person has an 0.7 chance, and a high-skilled person has an 0.9 chance. After working on the puzzles, the subjects are asked to rank themselves.

How should a fully rational, unbiased subject who manages to solve one puzzle correctly rank himself? Before attempting the puzzles, his beliefs about himself match the population distribution – there is a 55% chance he is low-skilled, a 25% chance he is medium-skilled and a 20% chance he is high-skilled. Now the distribution over his possible skill levels is:

$$\begin{aligned} p\left(\text{low skill} \mid \text{one correct}\right) &\approx 0.41 \\ p\left(\text{medium skill} \mid \text{one correct}\right) &\approx 0.44 \\ p\left(\text{high skill} \mid \text{one correct}\right) &= 0.15 \end{aligned}$$

There is about a 0.59 chance that his skill is either high or medium, i.e., is in the top 45% of the distribution. His expected probability of solving a sample puzzle is 0.48, which is better than for 55% of the population and better than the population mean of 0.41. His most likely type - medium - is better than the types of over half the population. Almost any way he looks at it, his best answer is that he ranks in the top half of the distribution. At the same time, if asked to reveal his beliefs by choosing between receiving a monetary prize with a 50% probability and receiving the prize if he is in the top half of the distribution, he should choose to bet on his placement. But, as Table 1 shows, 53% of the population will solve at least one sample puzzle correctly. Hence, 53% of subjects will (correctly) place themselves in the top half, or even the top 45%, giving the misleading appearance that the population is overconfident.

The effect is even more dramatic when we consider people who have answered two puzzles correctly. There is over a 0.5 chance that such a person's skill level is high, making high-skilled a reasonable self-placement. Since 29% of the population will solve both puzzles, 29% of the population can reasonably rank themselves in the top 20%.

Table 1. Posterior Beliefs over types, after 0,1 or 2 correct answers.

	Posteriors		
Type↓\# puzzles solved→	0	1	2
Low	0.95	0.41	0.02
Medium	0.05	0.44	0.42
High	0.00	0.15	0.56
Percent who solve #	47%	24%	29%

Note: Figures are rounded to two decimal places

[4] The fact that this distribution is not symmetric is not significant.

As this example shows, the *apparent* overconfidence of too many people placing themselves in an upper interval does not indicate *true* overconfidence. It is not difficult to understand why. Bayesian updating requires that the weighted sum of beliefs be correct. If a fraction x of the population believes that they rank in, say, the top half of the distribution with probability *at least* $q > \frac{1}{2}$, then Bayesian rationality immediately implies that $xq \leq \frac{1}{2}$, not that $x \leq \frac{1}{2}$. As Theorem 1 below shows, $xq \leq \frac{1}{2}$ is also sufficient for Bayesian rationality. Hence, while the particular numbers in this example resulted in 53% of the population placing themselves in the top half, different numbers could have yielded nearly 100% correctly believing that it is more than likely that they rank there. Put differently, if the experimenter does not know the true signalling structure from which the agents derive their beliefs, simply showing that a majority rate themselves as better than average is, at best, weak evidence of overconfidence. Several possibilities for a more meaningful test remain, including:

1. Have subjects place themselves into a narrower interval than the top 50% or ask that they place themselves with greater confidence than a 0.5 chance. For instance, a finding that 48% of subjects place themselves in the top 20% with probability at least 0.5 would be difficult to rationalize as Bayesian (see Theorem 1). Experiments along these lines stay within the spirit of the majority of prior better-than-average experiments. Indeed, some of those experiments already contain some of these elements. For instance, Svenson asks his subjects to place themselves into deciles, not merely above or below the median. We carry out these types of studies in Experiment I.

2. Elicit more complete information on the nature of subjects' beliefs. In the above example, the (unrounded) weighted sum of people's beliefs that they are in the top 45% is exactly 0.45, as it must be (see Theorem 3). If data were to yield a sum larger than 0.45, we could more readily conclude that the population is overconfident. We carry out this type of study in Experiment II. Clark and Friesen (2009), Moore and Healy (2008) and Merkle and Weber (2011) are also experiments that use more complete information.

3. Exploit information about subjects' actual abilities. In the above example, at least half of the 53% of the population that bet on themselves to place in the top 45% should turn out to actually be in the top 45% (see Theorem 2 below). We carry out this type of study in Experiments 1 and 2. Burks et al. (2013) also carry out a study that uses information on subjects' actual placements.

4. When a disproportionate number of people rank themselves high, they must be counterbalanced by a group of people who are relatively certain that they rank low. In the above example, the 53% of people who believe they are above average are offset by 47% who are quite certain (probability .95) that they are in the bottom 55%, and also quite certain (probability .86) that they rank in the bottom half if ties are broken randomly. A

useful test can be constructed by focussing on the beliefs of people who rank themselves low. We use such information in Experiment II. Merkle and Weber also (2011) exploit information on these below average beliefs.

5. Use information the experimenter may posses about the distribution of types in the population, and the signals agents are likely to have received, to draw conclusions. We do some analysis along these lines in Experiment II. (see Section 4.1 and also footnote 9).

Crucial to the possibility of rational overplacement as demonstrated in the above example, is the idea that people do not know their own types with certainty. If they were certain of their types, then only 50% could place themselves in the top half (absent ties). Within the behavioral economics literature, a number of papers, including Bénabou and Tirole (2002) and Kőszegi (2006) start from the premise that people are continually learning about their own types. Several strands of the psychology literature also stress that people are uncertain of their types, including Festinger's (1954) influential social comparison theory and Bem's (1967) self-perception theory.

These theories pit the motivation to obtain accurate information about the self against the motive to enjoy a positive self-image. The pursuit of a positive self-image will naturally contribute to overconfident beliefs (Kőszegi, 2006; Kunda, 1990). There may be rational reasons to hold false beliefs if those beliefs deliver utility. There are at least two ways in which they might do so. First, it is gratifying to hold a positive self-image (Baumeister, Campbell, Krueger, and Vohs, 2003). Second, in some domains self-confidence may contribute to greater success because it leads people to exert more effort or overcome risk aversion (Bénabou and Tirole, 2002; Kahneman and Lovallo, 1993).

Both of these causes for inaccurately positive self-assessments ought to depend on the perceived importance of the task or performance domain. For example, for most of us, it gives more pleasure to believe that we are good friends than to believe that we are good arborists. At the same time, the more important performance becomes, the more value there is in sustaining motivation to exert ourselves and overcome obstacles. In Experiment I, we vary the perceived importance of the task so that we can assess the power of motivated self-enhancement and the degree to which it affects better-than-average beliefs.

We also test whether subjects are indeed uncertain of their types. Note that such uncertainty creates an ambiguity when a subject is, in effect, asked for a point estimate of his type. A plausible interpretation of a subject's point answer is that it represents his beliefs about his median type, but other interpretations are also possible, including that the answer represents his mean or modal type. Clearly, it is desirable for the design of an experiment to restrict the possible interpretations of the answers, as the designs of some experiments do, including the ones we carry out here.

1.2 Relevant Literature and Results

To place this paper relative to the literature, consider the following chronology. B&D argued that most prior experiments on the better-than-average effect, while suggestive, did not conclusively demonstrate overconfidence, even accepting the experiments on their own terms.[5] Two papers that were relatively immune to the critique of B&D were Clark and Friesen (2009) and Moore and Healy (2008). Interestingly, in contrast to the bulk of the literature, these experiments did not find overconfidence.

In Clark and Friesen (2009), subjects solve sets of problems and are asked to predict how often their performance will place them in the top 5 out of a group of 12. In ten sessions, their aggregate prediction is indistinguishable from $\frac{5}{12}$, as it should be in a rational population. In one session, the aggregate prediction is below $\frac{5}{12}$ and in one it is above. In Moore and Healy (2008), subjects take eighteen ten-item quizzes and, prior to each quiz, are asked to give the probability of obtaining each possible score, both for themselves and for a randomly selected prior participant. The average across subjects of the expected value of their own score equals the average of the expected values for the random person, as it should. In both experiments, participants are rewarded using the quadratic scoring rule.

The first set of tests developed on the basis of B&D were conducted by us several years ago and are reported here as Experiment I. These tests were done in a manner that was comparable to that of prior studies, partly to help us assess whether or not that style of study is promising. Those tests did not find overconfidence but, as we will discuss, the tests were relatively weak and we decided to push further. More recently, we re-examined those same results using stronger tests and ran another experiment, reported here as Experiment II. Many of these newer tests find overconfidence. We check if this overconfidence can be explained by a motive for self-enhancement but fail to find support for this explanation.

Merkle and Weber (2011) and Burks et al. (2013) also conduct experiments that take into account the critique of B&D. In Merkle and Weber, individuals complete four tests on four different tasks – covering intelligence, memory, general knowledge, and creativity – and then predict the probabilities with which their score will fall into each decile. On each of the first three tasks, the total weight of belief that subjects place on being in the top half is greater than 50%, indicating overconfidence. On the fourth task, in contrast, there is no indication of overconfidence as subjects' beliefs are well calibrated. Subjects' payments are determined using the quadratic scoring rule.

In Burks et al. (2013), subjects take an I.Q. test and a numeracy test. Before each test, they are shown a single sample question and asked to predict the quintile in which their score will fall. They are paid $2 if their prediction matches the actual quintile. Subjects typically

[5] It is possible that the results of some of these experiments can be fortified by appealing to implicit assumptions the authors may have had in mind.

name quintiles greater than the one they end up falling into, which the authors characterize as overconfidence. The design of this experiment is novel, and Burks et al. use novel methods to analyze the data. However, while the methods they use are reasonable, other approaches to the same data can yield different conclusions. For instance, they adopt the Euclidean metric to measure the distances between samples and, although this metric is a natural one, other metrics yield different distances.[6]

In Eil and Rao (2011), groups of 10 individuals either complete an I.Q. test or are ranked by others according to their beauty as part of a speed-dating exercise. They then assess the probability with which they will fall into each of 10 ranking positions. The average expected placement they give is 5.2 for intelligence and 4.3 for beauty. Both figures are below the Bayesian prediction of 5.5. Eil and Rao do not conduct a statistical analysis of these results but, to us, it seems likely that 5.2 is indistinguishable from 5.5, while 4.3 shows underconfidence.

In none of these experiments, aside from the present ones, are subjects provided with information on the overall distribution of test scores. The theories in B&D and Moore and Healy (2008) both suggest that failure to do so may produce misleading evidence of overconfidence or underconfidence (see Section 5.2). In particular, the fact that Merkle and Weber find overconfidence only on their three easy (as it turns out) tasks is quite consistent with both these theories. Burks et al. model their subjects as receiving signals directly about their placements, rather than about their relative scores, which, formally, obviates the need for information on the distribution. The question, then, is whether this is an appropriate modelling in their context. The fact that their subjects are asked to predict their placement on unfamiliar quizzes, rather than self-evaluate on familiar tasks, suggests that it could have been especially important to provide them with a lot of information on the nature of the quizzes.[7]

In addition to theory, prior empirical studies, including Moore and Small (2007) and Pulford and Colman (1997), have found that it may be important to provide subjects with information on test scores. In contrast, the present study contains a manipulation that varies whether or not subjects are given explicit information on the distribution of test scores but

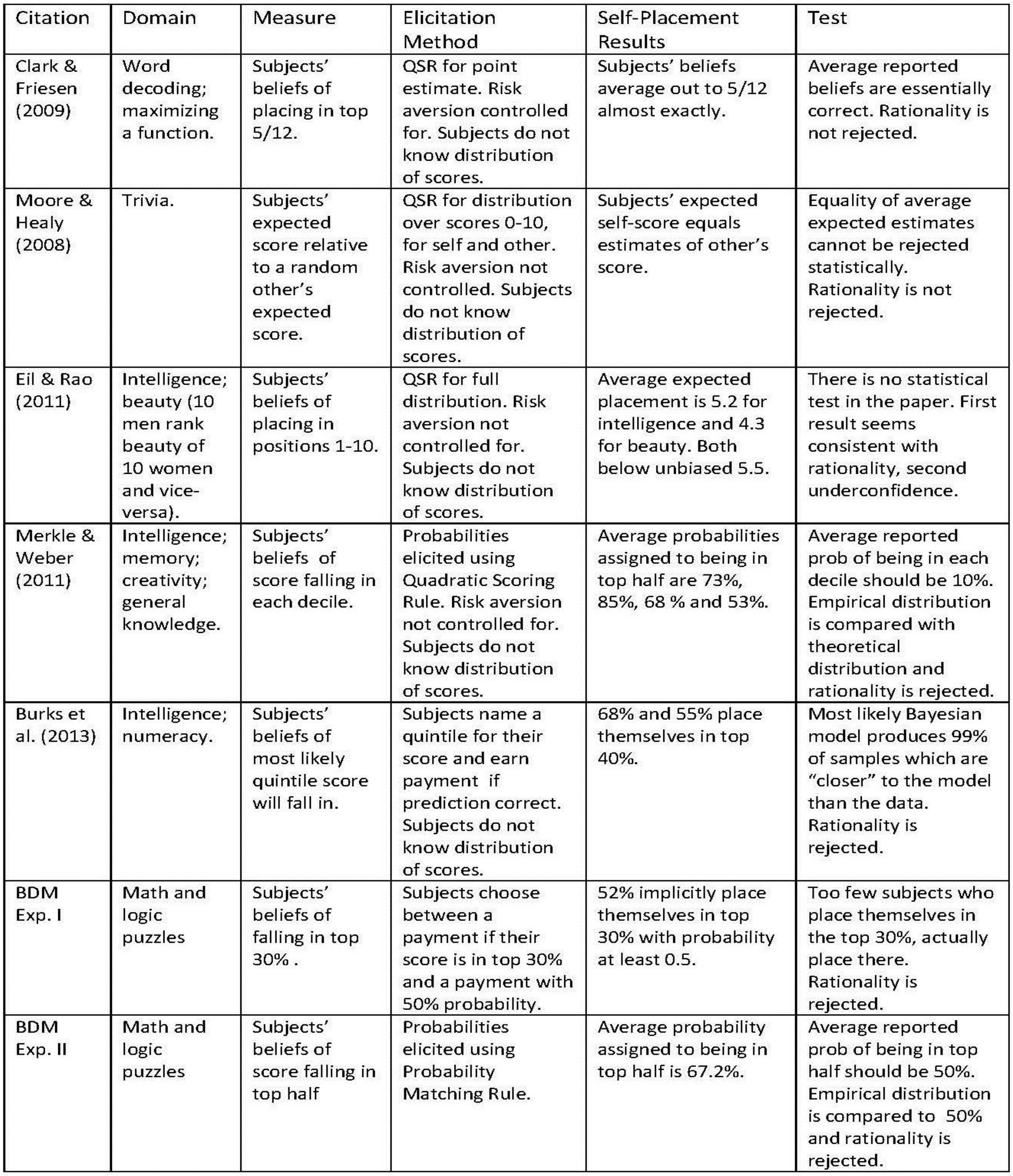
[6] In their model, the following three items are each 5x5 matrices whose entries, which are probabilities over types and declarations of likely quintiles, sum to one: i) the data they collect, ii) the rational model most likely to have generated the data, and iii) a sample of observations generated from the rational model. With the Euclidean metric, the distance of most samples to the model is smaller than the distance of the data to the model. However, different metrics yield different results. This is easiest to see with the discrete metric, where all samples are the same distance from the model. While the discrete metic is not an attractive one in this context, it is not clear what the statistical or theoretical justification is for using the Euclidean metric rather than, say, the max metric or another.

[7] Subjects in their experiment are truckers who, presumably, are not used to taking IQ and numeracy quizzes in their day-to-day lives.

does not find an effect. It is possible that we gave our student subjects sufficient information to gauge the difficulty of the test without the additional distributional information or that we happened to pick a test whose difficulty matched our subjects' priors (although we do not independently test the validity of these possibilities). It is also possible that information on difficulty is less important than theory and prior studies suggest. More research is called for here.

All these papers, including ours, are subject to several qualifications which we discuss in Section 5. Table 2 summarizes the findings of these papers.

Table 2. A summary of the most relevant papers.



Citation	Domain	Measure	Elicitation Method	Self-Placement Results	Test
Clark & Friesen (2009)	Word decoding; maximizing a function.	Subjects' beliefs of placing in top 5/12.	QSR for point estimate. Risk aversion controlled for. Subjects do not know distribution of scores.	Subjects' beliefs average out to 5/12 almost exactly.	Average reported beliefs are essentially correct. Rationality is not rejected.
Moore & Healy (2008)	Trivia.	Subjects' expected score relative to a random other's expected score.	QSR for distribution over scores 0-10, for self and other. Risk aversion not controlled. Subjects do not know distribution of scores.	Subjects' expected self-score equals estimates of other's score.	Equality of average expected estimates cannot be rejected statistically. Rationality is not rejected.
Eil & Rao (2011)	Intelligence; beauty (10 men rank beauty of 10 women and vice-versa).	Subjects' beliefs of placing in positions 1-10.	QSR for full distribution. Risk aversion not controlled for. Subjects do not know distribution of scores.	Average expected placement is 5.2 for intelligence and 4.3 for beauty. Both below unbiased 5.5.	There is no statistical test in the paper. First result seems consistent with rationality, second underconfidence.
Merkle & Weber (2011)	Intelligence; memory; creativity; general knowledge.	Subjects' beliefs of score falling in each decile.	Probabilities elicited using Quadratic Scoring Rule. Risk aversion not controlled for. Subjects do not know distribution of scores.	Average probabilities assigned to being in top half are 73%, 85%, 68 % and 53%.	Average reported prob of being in each decile should be 10%. Empirical distribution is compared with theoretical distribution and rationality is rejected.
Burks et al. (2013)	Intelligence; numeracy.	Subjects' beliefs of most likely quintile score will fall in.	Subjects name a quintile for their score and earn payment if prediction correct. Subjects do not know distribution of scores.	68% and 55% place themselves in top 40%.	Most likely Bayesian model produces 99% of samples which are "closer" to the model than the data. Rationality is rejected.
BDM Exp. I	Math and logic puzzles	Subjects' beliefs of falling in top 30% .	Subjects choose between a payment if their score is in top 30% and a payment with 50% probability.	52% implicitly place themselves in top 30% with probability at least 0.5.	Too few subjects who place themselves in the top 30%, actually place there. Rationality is rejected.
BDM Exp. II	Math and logic puzzles	Subjects' beliefs of score falling in top half	Probabilities elicited using Probability Matching Rule.	Average probability assigned to being in top half is 67.2%.	Average reported prob of being in top half should be 50%. Empirical distribution is compared to 50% and rationality is rejected.

Leaving aside the caveats and accepting the results of these papers at face value, a somewhat mixed picture emerges at this stage of research, with some experiments finding overconfidence and others finding well-calibrated subjects. To these papers, the reader should add whatever conclusions can be drawn from the long-standing research on the better-than-average effect, B&D's critique notwithstanding.

In the studies we report on here, we carry out a theory-based test as described in the next section.

2 Theoretical Underpinnings

B&D propose that data should not be called (prima facie) overconfident if it can be obtained from a population that derives its beliefs in a fully rational and consistent manner. To this end, they define a **rationalizing model** as a four-tuple $\left(\Theta, p, S, \{f_\theta\}_{\theta\in\Theta}\right)$, where $\Theta \subset \mathbf{R}$ is a finite type space, p is a prior probability distribution over Θ, S is a finite set of signals, and $\{f_\theta\}_{\theta\in\Theta}$ is a collection of likelihood functions: each f_θ is a probability distribution over S. The model has a straightforward interpretation. There is a large population of individuals. In period 0, nature draws a performance level, or type, for each individual independently from p. The prior p is common knowledge, but individuals are not informed directly of their own type. Rather, each agent receives information about himself from his personal experience. This information takes the form of a signal, with an individual of type $\theta \in \Theta$ receiving signal $s \in S$ with probability $f_\theta(s)$. Draws of signals are conditionally independent. Given his signal and the prior p, an agent updates his beliefs about his type using Bayes' rule whenever possible.

Data can be rationalized if it can arise from a population whose beliefs are generated within some rationalizing model.[8] Formally, fix $y \in (0,1)$ and let $\Theta_y = \left\{\theta : p\left(\widetilde{\theta} \geq \theta\right) \leq y\right\}$. Thus, Θ_y represents the top y of the population. For any $q \in (0,1)$, let $S_y^q = \{s \in S \mid p(\Theta_y \mid s) \geq q\}$; that is, S_y^q is the set of signals that cause a person to believe with probability at least q that his type is in the top y. For each s, let $f(s) = \sum_\Theta p(\theta) f_\theta(s)$; that is, f is the probability distribution over signals $s \in S$. Then, $f\left(S_y^q\right)$ is the (expected) fraction of people that believe that their type is in the top y with probability at least q.

- Let x be the fraction of the population that believes that there is a probability at least q that their types are in the top y of the population. We say that this data can be **rationalized** if there exists a rationalizing model $\left(\Theta, p, S, \{f_\theta\}_{\theta\in\Theta}\right)$ with $x = f\left(S_y^q\right)$.

[8] B&D use a more restrictive notion of rationalizability for their theorems – one where the analyst is not free to choose the type space and prior. The theorems in this paper are also true for this more attractive, but more difficult to state, notion.

- Let $\tilde{x}$ be the fraction of people that have beliefs as above and whose type is actually in the top y of the population. We say that this data can be rationalized if there exists a rationalizing model $\left(\Theta, p, S, \{f_\theta\}_{\theta\in\Theta}\right)$ with $x = f\left(S^q_y\right)$ and $\tilde{x} = \Pr\left\{(\theta, s) \in \Theta_y \times S^q_y\right\}$.
- Let r be the average belief that people have that their type is in the top y. We say that this data can be rationalized if there exists a rationalizing model $\left(\Theta, p, S, \{f_\theta\}_{\theta\in\Theta}\right)$ with $r = E_f\left[p\left(\Theta_y \mid s\right)\right]$

The basic idea behind rationalizing is that agents are not overconfident if they form their beliefs using the information available to them in an unbiased Bayesian manner. Ideally, we would like to compare the beliefs that agents have to the beliefs the agents would have if they formed their beliefs using the actual distribution of types in the world, all the (relevant) signals they have seen in their lifetime and the actual likelihoods of these signals. This *true* model is all but impossible to know, although in some circumstances some of its features may be discernible and usefully exploited.[9] The above definition of rationalizing is agnostic as to the nature of the true rationalizing model and calls data rationalizable if there is any model that rationalizes it. Data that is rationalizable could still come from agents who are overconfident or, for that matter, underconfident, relative to the beliefs they should have given their experience.[10]

While the definition of rationalizing allows for flexibility in proposing a rationalizing model, once one has been adopted, its requirements are stringent. Agents within the model have a perfect understanding of the model and all agree on the prior. In this respect, it is difficult to rationalize data. It is important to note that if, in practice, agents do not know the distribution of types in the population, then data that cannot be explained by a rationalizing model may still not be evidence of misconfidence (see Section 5.2). This caveat can be especially important in experiments where subjects are presented with somewhat unfamiliar tasks. For this reason, as we discuss later, we gave our subjects information on the distribution of types.

Without confronting the challenge of determining the true rationalizing model, we can still ask that the rationalizing model used to generate data be, in some sense, reasonable. B&D

[9] In particular, in some circumstances we might have information about the type of signals agents have seen or are likely to have seen. For instance, most people have met with suceess on easy tasks such as learning how to ride a bicycle, and failure on difficult tasks such as juggling three objects. B&D show that populations that are faced with easy tasks, on which their signals are predominantly successes, should display an apparent overconfidence, while populations that face difficult tasks should appear to be underconfidenct. This implication is confirmed by Moore and Healy (2008).

[10] Suppose, for instance, that half the population places themselves above the median and half below. Clearly, the data is rationalizable and even appears to be perfectly calibrated. Nonetheless, the population could be underconfident or overconfident relative to the beliefs they should have. For instance, if the example in the introduction represents the way the world actually works, then 53% of people should be placing themselves in the top half, so that the population is actually underconfdent. Indeed, if 52% of subjects placed themselves in the top half, they would still be underconfident.

propose one notion of reasonableness – that the likelihood functions satisfy the *monotone signal property* (akin to the monotone likelihood ratio property) – and we verify that our results our robust to this refinement.

We report on four different tests, which are based on the three theorems below. The theorems are essentially corollaries of theorems found in B&D, so their proofs are omitted.[11] Theorem 2 uses information about subjects' actual placement in addition to their beliefs. To the best of our knowledge, the first paper to use actual placement data to conduct a proper test of overplacement is Burks et al. (2013), although they use the placement data in a different way than we do, involving modal beliefs.

- Recall that a person of type θ is in the top y of a population if the fraction of people whose type is greater than or equal to θ is at most y. Thus, in a population of 100 people at most 25 can be in the top 0.25.

Theorem 1 *Suppose that a fraction x of the population believe that there is a probability at least q that their types are in the top $y < q$ of the population. These beliefs can be rationalized if and only if $xq \leq y$.*

Theorem 2 *Suppose that a fraction x of the population believe that there is a probability at least q that their types are in the top $y < q$ of the population. Let $\tilde{x}$ be the fraction of people who have those beliefs and whose actual type is in the top y of the population. This data can be rationalized if and only if $xq \leq \tilde{x}$.*

Theorem 3 *In a population of n individuals, let r_i, $i = 1, ..., n$, be the probability with which individual i believes his type is in the top y of the population. This data can be rationalized if and only if $\frac{1}{n}\sum_{i=1}^{n} r_i = y$.*

Because we are interested in overconfidence, Theorem 1 and Theorem 2 are stated only for the case $y < q$. Theorems for $y > q$ are essentially symmetric. We say that data **passes** a test based on one of the theorems if the necessary and sufficient condition in that theorem is satisfied and fails the test otherwise.

From Theorem 1, we can infer overconfidence if a sufficiently large fraction of people (variable x in the theorem) believe sufficiently strongly (variable q) that they rank sufficiently high (variable y). The most distinctive contribution of the theorem is represented in variable q, which quantifies the variation in how sure an agent is of her type. In our first experiment, 52% of subjects reveal themselves to be at least 50% sure they are in the top 30%. Here, $x = 0.52$, $q = 0.5$ and $y = 0.3$. Since $0.52 \times 0.50 = 0.26 < 0.30$, the data passes a test based on Theorem 1 and we cannot rule out the possibility that the agents are unbiased.

[11] The theorems are not exact corollaries because of slight discrepancies between the definitions used here and in B&D.

Theorem 2 is derived from the tautology that if agents have correct beliefs then those beliefs must be correct! For instance, in a large population, at least $\frac{3}{5}$ of agents who assign a probability of $\frac{3}{5}$ or greater to being above average, should actually be above average. To see that this is implied by the theorem, rewrite the necessary and sufficient condition as $q \leq \frac{\tilde{x}}{x}$. The term $\frac{\tilde{x}}{x}$ is the fraction of those people who believe they are in the top y that are, in fact, there. In our first experiment, while 52% of subjects believe it is likely they place in the top 30%, only 17% of subjects have this belief and rank there. Here, $x = 0.52$, $q = 0.5$ and $\tilde{x} = 0.17$. The data cannot be rationalized, since $0.52 \times 0.50 > 0.17$. Put differently, only $\frac{17}{52} = 33\%$ of subjects who believe they are at least 50% likely to place in the top half actually do.

Theorem 3 uses more detailed information about beliefs. Rational updating requires that the total weight that beliefs place in an interval, $\frac{1}{n}\sum_{i=1}^{n} r_i$, match the size of the interval. In our second experiment, the total weight that subjects place on being in the top half is 0.67. Here, $\frac{1}{n}\sum_{i=1}^{n} r_i = 0.67$ and $y = 0.5$. The data fails a Theorem 3 test since $0.67 \neq 0.5$.

Remark 1 *Suppose that an experiment yields data rich enough to perform tests based on all three theorems. If the data fails any single test, then the subjects in the experiment have beliefs that cannot be generated from a rationalizing model, regardless of whether or not the data passes other tests. It is trivial to see that if the data $(x, \tilde{x}, q, y)$ passes a test based on Theorem 2, it also passes a test based on Theorem 1 (since $\tilde{x} \leq y$). In a sense, then, a test based on Theorem 1 is made redundant by a test based on Theorem 2—if the data fails the latter test no further testing is required, while if the data passes the latter, a test based on the former will provide no relevant information. Similarly, if the data passes a test based on Theorem 3, it also passes one based on Theorem 1 (see Appendix B), so that the latter test is again seemingly made redundant. On the other hand, tests based on theorems 2 and 3 are independent of each other.*

Remark 2 *As remarked earlier, the model used to rationalize data may have little connection with the way agents should be forming their beliefs using the information available to them. Because of this, tests based on Theorem 1 may fail to find overconfidence that is revealed with tests based on Theorem 2. As an illustration, return to the puzzle-solving example from Section 1.1, but now suppose that, instead of updating correctly, the* 29% *of individuals who solve two puzzles correctly are overconfident – they believe they have an* 0.65 *chance of placing in the top* 20%*, rather than their actual* 0.59 *chance. A Theorem 2 test can detect this overconfidence, since only* 59% *of the people who believe that they rank in the top* 20% *with probability (at least)* 0.65 *will turn out to actually rank there. However, no Theorem 1 test will find a problem here (since* $0.29 \times 0.65 < 0.20$*). A Theorem 1 test can only conclude that these beliefs are rational for some population; it cannot detect that these beliefs are not rational for the population at hand.*

Remark 3 *Following up on the previous remark, note that data may fail a test based on Theorem 2 even if less than a fraction y places itself in the top y. For instance, if 5% of a (large) population places itself in the top 10%, but none of these 5% actually places there, the data will fail a Theorem 2 test. In this case, the failure is best viewed as indicating that the people who place themselves in the top* 10% *are overconfident, while the population on the whole is displaying a mixture of overconfidence and underconfidence. Also, if subjects misapprehend exactly what the test is testing for, then the "wrong" people may place themselves in the various deciles, even if their beliefs are correct with respect to their understanding of the nature of the test. A Theorem 2 test is adept at picking up on the fact that people are making errors relative to the beliefs they should have, though not necessarily at identifying the source of the errors.*

Remark 4 *Tests based on Theorems 1 and 2 are based on lower bounds of subjects' beliefs. Hence, they use relatively little information about subjects beliefs and may miss some overconfidence that is picked up by tests based on Theorem 3, when more complete information is available.*

Remarks 1, 2, and 4 suggest that Theorem 1 is of limited use. However, tests based on Theorem 1 have their own merit, as we discuss in Section 5.

We turn now to our experiments.

3 Experiment I

Subjects were recruited through the web site of the Center for Behavioral Decision Research at Carnegie Mellon University. The experiment was advertised under the name "Test yourself" with the description, "Participants in this study will take a test with logic and math puzzles. How much money people make depends on their performance and on how they choose to bet on that performance." This description, which suggests that participants' payments will depend on their skill, was chosen to be conducive to yielding overconfident-looking data (see Camerer and Lovallo (1999)), as we were interested in determining whether or not overconfident-looking data also indicates true overconfidence. Since prior experimental work and theory (see B&D and Moore and Healy (2008)) indicate that populations exhibit the better-than-average effect more markedly on easy tasks than difficult ones, the test that subjects took was designed to be easy.

We ended up with 134 subjects, 78 females and 56 males, with a mean age of 25 years (SD = 6.4). We report data for the 129 subjects who submitted complete responses to the three key choices we gave them. The results are unchanged when we analyze, for each question, all the answers we have for that question.

We start with an overview of the experiment. During the course of the experiment, subjects made three decisions, took a quiz, and then again made three decisions. Each decision consisted of choosing between one of two options: (i) a *skill option*, where the subject might earn $10, depending on his performance on the quiz, and (ii) a *gamble option*, where the subject might win $10 with known probability. (The terms “skill” and “gamble” were not used in the experiment; the names “Benchmark bet”, “Strength bet” and “High Placement bet” did not appear in the experiment, either). The precise options for each set of three decisions is shown in the table below.

Skill Option	Gamble Option
1. Benchmark bet. You will receive $10 if your test score puts you in the top half of previous test-takers. In other words, if your score is better than at least 50% of other test-takers, you will get $10. . .	1. Benchmark bet. There is a 50% chance you will receive $10. We have a bag with 5 blue poker chips and 5 red poker chips. You will reach in to the bag without looking and randomly select one of the poker chips. If the poker chip is blue, then you will get $10. If it is red, you will get nothing
2. High Placement bet. You will receive $10 if your test score puts you in the top 30% of previous test-takers. In other words, if your score is better than at least 70% of other test takers, you will get $10. . .	2. High Placement bet. There is a 50% chance you will receive $10. We have a bag with 5 blue poker chips and 5 red poker chips. You will reach in to the bag without looking and randomly select one of the poker chips. If the poker chip is blue, then you will get $10. If it is red, you will get nothing.
3. Strength bet. You will receive $10 if your test score puts you in the top half of previous test-takers. In other words, if your score is better than at least 50% of other test takers, you will get $10 . .	3. Strength bet. There is a 60% chance you will receive $10. We have a bag with 6 blue poker chips and 4 red poker chips. You will reach in to the bag without looking and randomly select one of the poker chips. If the poker chip is blue, then you will get $10. If it is red, you will get nothing.

The Benchmark bet parallels the tests used in much of the research on the better-than-average effect. Choosing the skill option is strictly optimal for a subject who is more than 50% sure that his or her quiz performance will rank in the top half. Since it is possible for 99% of rational Bayesian subjects to have such a belief, the set of choices here does not, by itself, test for overconfidence. However, it facilitates a direct comparison of our results with prior studies.

The High Placement bet tests whether subjects expect to do particularly well. The skill option is attractive for those who believe they have more than a 50% chance of placing in the top 30%. Recall that, in Svenson’s unincentivized experiment (where measuring performance would have been difficult), over 80% of American subjects place themselves in the top 30% of drivers.

The Strength bet tests the strength (or subjective likelihood) of the same belief tested by the Benchmark bet. The skill option again pays off for a placement in the top half, but the gamble option is now more attractive, offering a 60% chance of winning. We use 60% because, independently of overconfidence, we are interested in whether a relatively small change in the chance of receiving a prize randomly — from 50% in the benchmark condition to 60% here – makes a significant number of subjects change their choices. That is, we are interested in the extent to which people are uncertain of their types.

A detailed description of the experiment follows.

1. Subjects were randomly assigned to experimental conditions that crossed two treatment variables, motivation (high or low) and knowledge of others' prior performance (yes or no). The motivation manipulation varied what subjects were told about the quiz they were about to take. With this manipulation we hoped to observe the effect of inducing a motive to be overconfident. The prior performance manipulation varied whether or not subjects were told how populations similar to theirs had performed on the quiz in the past. Those who received this information should have had enough information to form accurate priors about the likely distribution of scores.

2. Subjects in the high motivation condition were given this text to read:

   "In this experiment, you will be taking an intelligence test. Intelligence, as you know, is an important dimension on which people differ. There are many positive things associated with higher intelligence, including the fact that more intelligent people are more likely to get better grades and advance farther in their schooling. It may not be surprising to you that more intelligent people also tend to earn more money professionally. Indeed, according to research by Beaton (1975) ten I.Q. points are worth about four thousand dollars in annual salary. Children's intelligence is a good predictor of their future economic success according to Herrnstein and Murray (1994). Of course, this is partly because, as documented in research by Lord, DeVader, and Alliger (1986) intelligent people are perceived to have greater leadership potential and are given greater professional opportunities. But what may be surprising to you is that intelligent people also tend to have significantly better health and longer life expectancies (see research by Gottfredson and Deary, 2004)."

   Subjects in the low motivation condition read:

   "In this experiment, you will be taking a test of math and logic puzzles."

3. Subjects were given a set of sample quiz items. In order to constitute this set of sample items, we began with a larger set of 40 quiz items. One half of this set was randomly

chosen for Test Set S. The other half belonged to Test Set M. Those participants who were to take Test S saw sample items from Set M, and vice versa.

4. Subjects in the prior performance condition received a histogram showing how previous individuals had scored on the quiz they were about to take.

5. Subjects chose between skill and gamble options for each of the three previously described conditions. The order in which the three choices appeared was varied randomly, as was whether the chance or the skill option appeared first. Participants were told that they would make the three choices again after taking the test, and that one of these six choices would be randomly selected at the end of the experiment to count for actual payoffs. The results we present are those of the first set of choices; those made before taking the test.[12]

6. Subjects took the twenty-item test under a ten-minute time limit. A sample of the two test sets appears in Appendix A. Subjects earned $.25 for each test question they answered correctly.

7. Subjects again chose between the skill and chance options for each of the three conditions.

8. One of the six decisions a subject made was randomly chosen to deterimine his or her payoff. If a subject chose the gamble option rather than the skill option for the one choice that counted for payoffs, the subject drew from the relevant bag of poker chips to determine whether he or she won the $10 prize.

Afterwards, subjects answered a series of unincentivized questions regarding what they thought their score would be, how they felt during the experiment, and their motivation to perform.

3.1 The data

On average, our subjects answered 17.3 out of 20 of the test items correctly.

None of our five between-subjects experimental variations significantly affected subjects' choices (or their scores—except for the High Motivation treatment, which decreased scores; see Section 3.1.1). Firstly, as expected, there was no effect from any of the three randomizations of order, namely, the order of the presentation of the bets (123, 132, 213, etc.), whether the skill or random bet was presented first in each pair, and whether subjects saw sample M

[12] This is the standard methodology for studying overplacement (see Moore and Healy, 2008; Clark and Friesen, 2009; Hoelzl and Rustichini, 2005; inter alia). The second set of bets is more informative about how good subjects are at estimating their own scores after the fact. We do not present these data, but they are available online at: http://learnmoore.org/mooredata/OJD/

and took test S, or saw S and took M. Secondly, and less expected, the prior performance manipulation did not affect scores or choices between bets. Finally, and surprisingly to us, the Motivation manipulation had no effect either. Hence, we now discuss only aggregate data, without discriminating by between-subjects treatments.

Of paramount importance to a subject is her score on the test. Thus, it is most convenient to model a subject's *type* as just being this score. This means that at the time she makes her decision, the subject does not yet have a type. Rather, her type is a random variable to be determined later. Formally, this poses no difficulties. Based on her life experiences and the sample test she sees, the subject has a distribution over her possible types, i.e., test scores. On the Benchmark bet, a subject (presumably) prefers to be rewarded based on her placement if there is more than a 50% chance her type is in the top 50%. On the Strength bet, a subject prefers to be rewarded based on her placement if there is more than a 60% chance that her type is in the top 50%. On the High Placement bet, a subject prefers to be rewarded based on her placement if there is more than a 50% chance her type is in the top 30%.

Table 3 summarizes our findings.

Table 3. Choices for the three pairs of bets for Experiment I

	(x)	$\left(\frac{y}{q}\right)$	$(\widetilde{x})$	(xq)	p-value
	% who	max. allowed	% choose Skill &	min. required	See
	choose Skill	by Theorem 1	place in top y	by Theorem 2	Appendix C
Benchmark: 50% belief in top half $q = 0.5, y = 0.5$	74, data passes according to Theorem 1	100	39, data passes according to Theorem 2	37	
Strength: 60% belief in top half $q = 0.6, y = 0.5$	64, data passes according to Theorem 1	83	35, data fails according to Theorem 2	39	17% "passes" statistically
High: 50% belief in top 30% $q = 0.5, y = 0.3$	52, data passes according to Theorem 1	60	17, data fails according to Theorem 2	26	$< 1\%$ fails statistically

Notes: x is the percentage who choose the Skill bet; q is a lower bound on their beliefs they will place in the top y of scores; $\tilde{x}$ is the percentage of those who declare to be in the top y with probability at least q who actually place in the top y of scores.

As expected, on the Benchmark bet, the population displays apparent overplacement: 74% choose to be rewarded based upon their placement. Such a result is usually loosely interpreted as 74% place themselves in the top half of test takers. However, a more precise interpretation is that 74% believe that there is at least a 50% chance that they are in the top half.

Note that these two interpretations may have different implications for rationality. With the first interpretation, if we assume "place themselves" indicates (near) certainty, then the population displays overconfidence, not just apparent overconfidence. But the more precise interpretation, the second interpretation, shows that the choice behavior of the subjects is consistent with rationality, as indicated by Theorem 1.

Before turning to the question of overplacement, we consider the question of how certain a subject is of her type. Of the 74% who opt for placing in the top half over a 50% chance draw on the Benchmark bet, 22% switch and choose a 60% chance draw over placing in the top half on the Strength bet.[13] Thus, a significant fraction of the subjects do not show much confidence in their belief that they are better than average. This fact supports the underlying premise of B&D and of Moore and Healy (2008), that people are uncertain of their types.[14]

We turn now to the question of overplacement.

Tests Based on Theorem 1.

According to Theorem 1, the population exhibits overconfidence if more than 83% choose the skill option from the Strength bet or more than 60% choose the skill option from the High Placement bet. In the Strength bet, only 64% choose the skill option, so that rationality cannot be rejected. More precisely, one can (easily) build a rational model in which a sample at least this overconfident looking arises with probability greater than 50%, so that a null of rationality cannot be rejected. Similarly, the figure of 52% who choose the skill option on the High Placement bet is well below the threshold of 60% and is consistent with rational beliefs.

Further tests can be constructed by combining information gleaned from the different bets. In particular, the fact that 64% choose the skill option from the Strength bet restricts the number who can choose the skill option in the Benchmark bet. The results can also be tested against the requirement that the rationalizing model used be a reasonable one, in the sense of satisfying the monotone signal property from B&D. We perform these extra tests, and the data passes these as well (details are available from the authors). Despite all these successes, clearing Theorem 1's conservative hurdle for conclusive evidence of biased beliefs does not rule out overconfidence. The fact that we have information on actual test performance allows us to conduct more sensitive tests based on Theorem 2.

Tests Based on Theorem 2.

Before turning to the data, note that our theorems provide conditions that the data must satisfy when it comes from an arbitrarily large rational population. Of course, our data comes from a finite population. To test whether data from our experiment that fails a Theorem 2 test can nonetheless be said to pass in a statistical sense, we specify a rationalizing model and check how likely it is that sampling 129 subjects from a large population could have resulted in an outcome at least as overconfident as the data. Since our null hypothesis is rationality,

[13] We note that 6% of the subjects favor a 50% draw over their placement, but their placement over a 60% draw. We have no explanation for this inconsistent behaviour.

[14] Our experiment does not provide a definitive test of subjects' uncertainty about their types as they may also have been concerned about randomness in the test itself (although concern about this randomness should have been low, since subjects were shown a representative sample test).

we use the rationalizing model that has the best chance of generating such an outcome. It turns out that the probability the best-chance rationalizing model generates an outcome at least as overconfident as the data $(n, q, y, x, \widetilde{x})$ is given by $P(w \leq n\widetilde{x})$, where w is a random variable with binomial distribution $B(nx, q)$ (see Appendix C for details).

We have the following results:

1. In the Benchmark bet, 74% of the subjects choose the skill option. From Theorem 2, 37% of the subjects should have scores that in addition place them in the top half. Since 39% of subjects have scores also in the top half,[15] the data passes the rationality test.

2. In the Strength bet, 64% of the subjects choose the skill option. From Theorem 2, 39% should also place in the top half. In fact, only 35% do. However, while 35 is less than 39, there is a 17% chance that a sample as apparently overconfident as this, or more, can arise from a rational population, since $P(w \leq 129 \times .35) = 17$, for $w \sim B(129 \times 0.64, 0.6)$. Hence, we do not reject rationality.

3. In the High Placement bet, 52% choose the skill option. From Theorem 2, at least 26% should also place in the top half. In fact, only 17% do. The chance of this arising from a rational population is less than 1%, since $P(w \leq 129 \times 0.17) < 0.01$ for $w \sim B(129 \times 0.52, 0.5)$. Hence, we reject rationality.

Combining the results from tests based on Theorems 1 and 2, the data passes four out of five tests. However, to be deemed rational the data must pass all the tests.[16] Thus, the results of Experiment I reject the hypothesis that subjects are behaving rationally.

3.1.1 Motivation and prior performance manipulations

It is notable that the motivation treatment does not appear to increase overconfident beliefs. Those in the high motivation treatment choose the skill option 63% of the time, whereas those in the low motivation treatment do so 64% of the time. While the high/low motivation does not affect the betting behaviour of our subjects, they have significantly lower scores under the high motivation treatment. Those in the high motivation condition answer an average of 16.6 questions correctly, those in the low motivation condition answer an average of 18 correctly, and an independent samples t-test reveals this difference to be significant at significance levels below 1%. Thus, our subjects appear to "choke" under pressure, as others

[15] In determing subjects' placement, it was sometimes necessary to break tie scores randomly. Subjects knew this procedure would be used.

[16] Our results do not seem to be subject to the well-known multiple testing problem, whereby one must adjust the p-values for the fact that one is running several tests. In any case, our data fails one of only three tests based on Theorem 2, and its p-value is less than 1%.

have found (Ariely, Gneezy, Loewenstein, and Mazar, 2005; Beilock and Carr, 2001; Dohmen, 2005; Markman and Maddox, 2006). This finding speaks to the potential adaptiveness (or lack thereof) of motivations to perform. Logg, Moore, and Haran (2013) re-analyze this same data, focusing on the failure of the high-motivation manipulation to enhance either performance or overconfidence. They explore the implications of this result for theories that have attempted to explain overconfidence as a motivated bias.

The results are also robust to variation in prior performance information. Subjects in the prior performance treatment select the skill bet 66% of the time and 61% of the time in the no-prior performance treatment. A Chi-square test fails to reject the null hypothesis that these proportions are not significantly different from one another, Chi-square (1) = .645. The lack of an effect of this manipulation contrasts with prior evidence suggesting the powerful role that expectations of task difficulty have on estimates of performance (Moore and Small, 2007; Pulford and Colman, 1997). A possible explanation for this (non-) result is that providing our subjects with additional detail about quiz scores has little effect here because subjects began with fairly accurate expectations about how difficult the quiz would be. It is also possible that subjects pay insufficient attention to the population distribution of scores. We return to this last possibility in Section 5.3.

4 Experiment II

In this section, we report on a second experiment that allows for a test based on Theorem 3, as well as tests based on the first two theorems. The experiment is similar in its overall design to Experiment I. It again involves Carnegie Mellon undergraduates, 74 this time, taking a quiz similar to the previous ones. The crucial difference in the experiment is that we ask subjects, in an incentive compatible manner, to indicate the likelihood they ascribe to placing in the top half. The elicitation mechanism used is the probability matching rule described by Karni (2009) and Grether (1981), as implemented by steps 3, 4 and 7 below. The experiment proceeded as follows:

1. Participants took a five-item practice quiz. They had 2.5 minutes.

2. Participants were informed that the median score on previous tests was 18.

3. The experimenter described the probability matching rule and its incentive properties.

4. Participants indicated how likely they thought it would be that they would rank in the top half of quiz takers by choosing a probability from a drop-down menu. The menu listed the probabilities from 0% to 100% in 2% increments. Because of the nature of the interface, the menu had a probability on which it started – this probability was randomly determined for each participant.

5. Participants who indicated an 86% or larger probability of scoring in the top half were presented the following additional bet: Choose between the following two options, a) Lose \$1 if your score is not in the top half, or b) Lose \$1 with a chance of 20%. Participants did not know beforehand that this extra bet would be proposed.

6. After these choices, subjects took the twenty item quiz. They had 10 minutes.

7. The computer chose an even number uniformly from 0% to 100%. Participants who had indicated a number equal to or larger than the computer's number were paid \$10 if their score was in the top half. Participants who had chosen a number smaller that the computer's, drew a bingo ball from a cage with integers from 1 to 100; if the number on the ball was smaller than the number chosen by the computer, they won \$10.

With the probability matching rule, it is optimal for expected utility maximizing subjects that care only about money to report their true subjective probabilities when they can choose any number from the interval [0, 100]. When restricted to choosing even numbers, it is optimal for these subjects to round their beliefs to the nearest even number.

The reason for Step 5 is that we wanted to make sure that people who chose a high probability "really meant it." Therefore, we checked if participants who indicated a probability above 84% would act consistently with this estimate when presented with another bet that implied at least an 80% chance of ending in the top half. Of the fifteen people who indicated a probability above 84%, thirteen followed up in a consistent manner by choosing 5a over 5b.

From Theorem 3, the average of the likelihoods of ending in the top half given by participants should be 50% in a rational population, although given the restriction to even numbers and the rounding noted above, this figure could rationally be as high as 51% in the experiment. The actual average given was 67%, which is greater than 51% at all conventional confidence levels: the t statistic with 73 degrees of freedom is 7.06, which yields a p value of less than 1%. Thus, this test rejects the hypothesis that subjects were behaving rationally.

Tests based on theorems 1 and 2.

The data also enable us to conduct an additional nineteen tests based on Theorem 1 and sixteen based on Theorem 2.

- For instance, 35% of subjects indicate they have a probability of at least 0.8, or 0.79 when we allow for rounding, of ending in the top half. From Theorem 1, up to 62% of subjects could rationally make such an indication, so the data pass this test.
    - Of these subjects, 58% are actually in the top half. This data does not pass the test, as there is less that a 1% chance that a sample as apparently overconfident as this, or more, will arise from a rational population – $P\left(w \leq 74 \times .58\right) < 0.01$, for $w \sim B\left(74 \times 0.35, 0.79\right)$.

In this particular case the data passes the first test and fails the second test.

A complete list of the tests is provided in the appendix. Although the test based on Theorem 3 indicates that the beliefs from Experiment II cannot be rationalized, the data passes every test based on Theorem 1. At the same time, the data fails six tests based on Theorem 2 at the 5% confidence level, and fail eight tests at the 10% confidence level. This is consistent with the results of Experiment I, where tests based on Theorem 1 were not capable of detecting overconfidence.

4.1 Updating from signals

We know very little about the signals that participants received prior to coming to the experiment. We do know, however, that during the experiment they received an additional signal from the five sample questions. Presumably, the ability to answer more questions correctly is a positive signal.

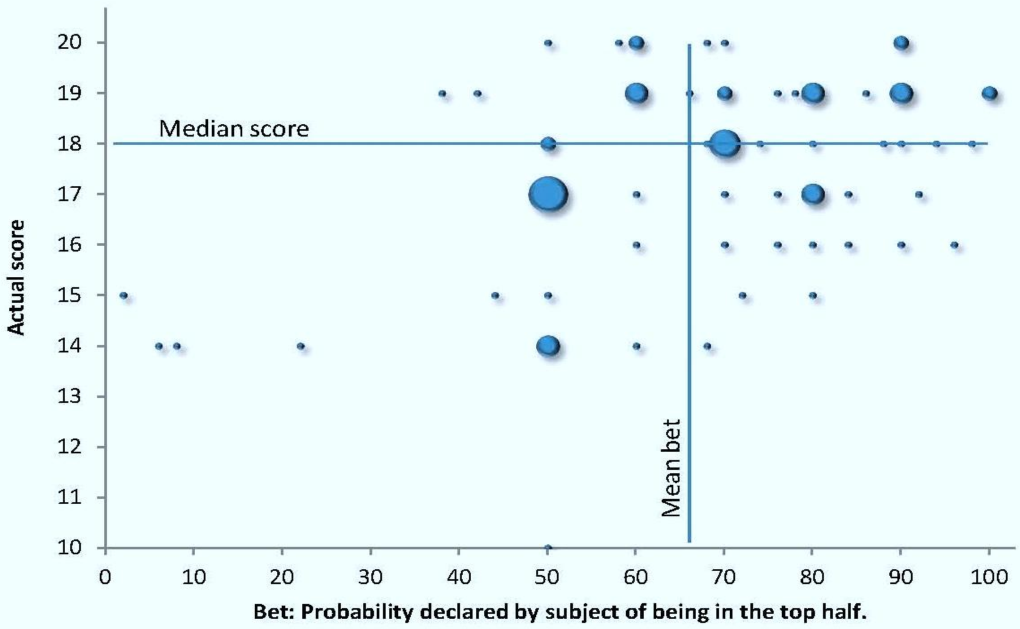
Table 4 reports performance information on the sample questions. The subjects were not incentivized to answer the sample questions correctly, so that the number of correct answers they provided is an imperfect measure of the number they could have provided. Indeed, it is quite likely that the three people who did not answer any sample questions correctly were simply not trying to. Nonetheless, as Table 4 shows, answering all five questions correctly is a strong predictor of performance: 71% of subjects who answer all five questions correctly place in the top half, a figure significantly greater than the percentages for those who answer fewer than five correctly. Despite this, subjects who answer three or four questions correctly report nearly the same probability of placing in the top half as those that answer all five correctly (very few subjects manage fewer than 3 correct answers). This suggests that subjects were not updating in a rational manner. Interestingly, subjects who answer five questions correctly indicate a 71% probability on average of placing in the top half, which is exactly correct. The overconfidence stems from those who answer fewer than five correctly. This is consistent with the hypothesis of Kruger and Dunning (1999) that overconfidence in the population is due to people who are, as they put it, "unskilled and unaware."

Table 4. Percentage of subjects placing in top half, depending on sample score

Score on Sample →	0	1	2	3	4	5
Number of subjects with score	3	1	3	14	22	31
Percentage placing in top half	33	0	0	36	41	71
Average reported probability of top half	55	60	26	68	69	71

As for their scores on the actual test, Figure 1 shows that there is a weak correlation between subjects' bets and their actual scores.

Figure 1. Bets and actual scores in Experiment II.
Dot size proportional to the number of subjects it represents.



5 Discussion

As indicated in Remark 1, tests based on Theorem 1 are weaker than tests based on Theorem 2 or Theorem 3. As indicated in Remark 2, there are manifestations of overconfidence that can be picked up by tests based on Theorem 2 but not by tests based on Theorem 1. Consistent with these two remarks, our tests based on Theorem 1 fail to find overconfidence in our experiments, whereas tests based on the other two theorems do find overconfidence.

Although tests based on Theorem 1 are formally weaker, three factors make them valuable. First, these tests can be applied to pre-existing experiments for which the data needed for tests based on the other theorems is not available. For instance, Svenson (1981) finds that 82.5% of American subjects in his experiment claim to be in the top 30% of drivers by skill level. Theorem 1 allows us to conclude that these drivers display overconfidence, as more than 60% believe that they are in the top 30% (if we assume that subjects rank themselves by their beliefs about their median types). Second, there may be reasons to have more confidence, at a practical level, in data elicited for tests based on Theorem 1 than in the more detailed data elicited for tests based on Theorem 3. Specifically, there may be a trade-off between the amount of information collected and the reliability of the information. Third, there are theoretical issues that may favour tests based on Theorem 1 over those based on Theorem 2. We now discuss points two and three.

5.1 Critical Assessment of Data

We took pains to provide incentives for accurate reporting. Despite this, there are at least two ways in which our data could be questioned.

1. Subjects probably had goals beyond maximizing monetary payments. In addition to the benefits of savoring positive self-regard from positive self-statements, people like to exert control over their situations, and so subjects may have preferred to bet on themselves even if they thought their chance of doing well was relatively poor (see Heath and Tversky, 1991; Goodie, 2003; Goodie and Young, 2007; and the references therein). While any such motivations could be thought of as introducing a sort of bias, the bias is distinct from overconfidence. These motivations would have to compete with losses in payment from suboptimal betting. How large are these losses?

   Subjects stood to gain \$10 from winning a bet. While this is a respectable wage for a half an hour's work, it (inevitably) overstates the incentives. For example, in Experiment I, a subject who chooses the skill option on the Strength bet even though she believes she has only a 34% chance of finishing in the top half, thus implicitly overstating her probability of success by 26%, suffers an expected loss of \$2.60, not \$10, from this suboptimal choice.[17] In Experiment II, a subject who overstates her probability of finishing in the top half by 26% suffers an expected loss of just \$0.33 (see Appendix B). By this accounting, the overconfidence in both experiments may be overstated, and the data from Experiment II may be more susceptible than the data from Experiment I. Similar caveats apply to the data from Burks et al. (2013). In their experiment, a subject earns \$2 for correctly predicting the quintile into which he or she will land. A subject's expected loss in payment from stating a higher quintile than his or her best guess may be quite small. Merkle and Weber (2011) do not report the exact formula they use to reward their subjects, so we cannot estimate subjects' losses from overplacing but it is a fairly general phenomenon that the expected losses from overstating a probability will be small compared to the prize at hand.

2. Although we carefully explained to subjects in Experiment II that declaring their true values was a dominant strategy, the argument is a bit subtle and it is possible that subjects did not understand it.[18] For instance, some subjects may have erroneously reasoned that, since stating a higher value ensures that when the randomizing device is

[17] In fact, subjects placed six bets and were rewarded based on a bet chosen at random. The calculation of \$2.6 assumes that subjects consistently overstate across the bets. Overstating only on some bets reduces the expected loss.

[18] Teachers of auction theory know that the simpler proposition that bidding one's value is a dominant strategy in a second price private value auction is far from obvious to most students.

used it has a higher probability of paying off, it is desirable to overstate.[19] Similarly, Merkle and Weber (2011) and others use the rather unintuitive Quadratic Scoring Rule to elicit beliefs.

The above two points suggest that our findings of overconfidence may be overstated, and more so for Experiment II than Experiment I. Since our two experiments are similar with regard to the subject population and the quizzes involved, it makes sense to compare the results from the two to see if there is evidence that subjects are overplacing themselves in Experiment II relative to the Experiment I.

In Experiment I, 74% of subjects bet on themselves to place in the top half. If we assume that subjects are straightforward expected utility maximizers, then 74% of the population believe they have at least an 0.5 probability of placing in the top half. In Experiment II, 91% of subjects report at least an 0.5 chance of placing in the top half, which is significantly greater than 74%. By themselves, these numbers suggests that the above concerns are warranted. However, the raw numbers are a bit misleading. In Experiment II, 18% of subjects report a probability of exactly 0.5. If we make a genericity assumption and assume that half of these were the result of rounding up and half of rounding down, then we have 82% with a belief greater than 0.5. We then cannot reject the hypothesis that the samples from the two experiments have the same mean: the t statistic for 2 samples with unequal variances is 1.11, and has 165 degrees of freedom; the p-value is 26.7%.

In Experiment I, 64% indicate a belief of at least 60% that they place in the top half. In Experiment II, 72% indicate a probability of at least 60%. Again, if we exclude half of those who say exactly 60% on the grounds that they have rounded up, then the relevant figure is 66%, and we cannot reject the hypothesis that the two samples have the same mean: the t statistic for 2 samples with unequal variances is 0.26, and has 153 degrees of freedom; the p-value is 78.8%.

Thus, there is some evidence that the mechanism used in Experiment II did not cause participants to overstate their placement relative to the mechanism used in Experiment I, which, in turn, provides some evidence that the above two concerns are not important. However, this evidence is not conclusive.

5.2 A Reassessment of the Theory

A quiz provides a clearer measure of performance than can easily be obtained on, say, driving or managerial skills. However, a quiz-based experiment suffers from the fact that subjects must reflect not only upon their skills but also upon the nature of the quiz they are taking.

[19] See also Plott and Zeiler (2005), whose results show that some of the findings confirming the endowment effect may have been the result of poor training by subjects on the Becker-DeGroot mechanism, which is the basis of the probability matching method we use.

Moore and Healy (2008) show that when subjects face a quiz that is easier than they expected it to be, even Bayesian updating may result in data that cannot be rationalized. The reason is that a subject who does well on the sample questions will be uncertain if this is because he is particularly skilled at this type of quiz or because the quiz is easy (so that many people will do well). He will rationally put weight on both possibilities and if the quiz is, in fact, easy, he will have placed too much weight on his skill (ex post). More generally, if subjects are uncertain of the actual distribution of scores, the data may misleadingly seem truly overconfident (see B&D (2011) for a discussion). To address this issue, in most treatments across our two experiments we gave subjects information about how people score . In the one treatment where we withheld information, subjects had information about the task in the form of 20 sample questions. Subjects in Burks et al. (2013) seem to have been given only a single sample question. Subjects in Merkle and Weber (2011) were asked for estimates after taking the quiz, so that they had significant information about the task. However, they did not have information on the performance of others.

There is another issue, which manifests itself when applying Theorem 2. Subjects must consider not only the ease of the quiz, but its diagnostic value as well (recall Remark 3). If subjects go into the quiz thinking it is on inductive reasoning, but it is turns out to be on deductive reasoning, and skills in these two domains are imperfectly correlated, data might misleadingly fail a test based on Theorem 2. Such a test is correctly picking up on the fact that subjects have made an error, but the error is one of misunderstanding the nature of the quiz, not one of overconfidence. The methodology used in Burks et al (2013) suffers from similar potential flaws. Indeed, their findings may be particularly susceptible to misinterpretation, as subjects simply pick their most likely quintile without giving any indication of how confident they are that they will land there.

Thus far, we have described an “error” on the part of the subjects, who do not properly understand the nature of the quizzes they face. However, it may instead be the analyst who is making a mistake. Suppose that all subjects correctly understand that some quizzes are more diagnostically valid than others. Moreover, they use the actual distribution of quiz types in the world in making their Bayesian calculations. These subjects are perfectly rational and understand the differing nature of quizzes perfectly, although they have imperfect information about the particular quiz they are taking. Correctly averaging data over all populations taking all quizzes, the data will pass a test based on Theorem 2. However, the experimenter — in the present case us — is applying the test to this particular experiment, and is not averaging across all experiments. The data may again fail the test, although now it is the analyst who is making an error, not the subjects.

5.3 Why overconfidence?

Leaving the caveats aside and accepting the results, what can account for the overplacement we find? The data is not rationalizable in the sense of B&D. Since we provide our subjects good information about task difficulty and the population of scores against which they will be compared, Moore and Healy's (2008) theory does not predict overplacement. Hence, the data seems to display (true) overconfidence.[20]

The motivation to believe in oneself and the subjective utility of inflated self-views (Bénabou and Tirole, 2002) does a poor job accounting for this overconfidence, given the failure of the motivation manipulation in Experiment I. Our failure to find that task importance contributed to better-than-average beliefs in Experiment I is noteworthy, given the plethora of evidence showing a correlation between perceived importance and overplacement (Dunning, 2005). However, there are other compelling explanations for those correlational results, including the possibility that people choose to develop those skills they regard as important (van den Steen, 2004). Our results do not do anything to bolster the case for a motivational cause of overconfident beliefs

Given the poor showing of motivation as an explanation for overplacement, what can account for our results? One viable explanation is a sort of myopia in comparative judgment, in which people conflate absolute evaluation with relative evaluation (Klar and Giladi, 1999; Windschitl, Rose, Stalfleet, and Smith, 2008), and wind up believing that if they perform well, they must be above average. This theory is consistent with our finding that prior performance information had little effect on the results. After all, if subjects only take into account their own performance, information on others is irrelevant. This theory also comports well with the preponderance of evidence showing that on hard tasks, for rare events, and when failure is more common than success, people tend to believe that they are worse than others (Moore, 2007).

6 Economic Implications

Let us compare three situations. In the first, 80% of people rank themselves above the median but this is fully justified – there is only apparent overconfidence. In the second, 80% of people rank themselves above the median and this is not justified; rather, it is the result of overconfidence. In the third, in an experimental setting, 80% of people bet on themselves to place in the top half and this is not justified by objective evidence, but neither is it because they are overconfident; rather, it is because they like to exert control by betting on themselves.

[20] It is possible that above average subjects self-selected into the experiments but failed to correct for this self-selection. In that case, subjects display overconfidence with respect to their placement in the self-selected group. They may or may not be overconfident with respect to their placement in the overall population.

The observed outcome is the same in all three cases, so do the underlying differences really matter? We argue that the answer is 'yes'.

First take the distinction between apparent overconfidence and true overconfidence. Consider the question of whether authorities should regulate the behaviour of drivers by imposing speed limits, mandating seat belt use, etc..., rather than simply informing drivers of the risks. *One* argument in favour of such regulations is that drivers have too much confidence in their abilities. As Svenson writes, "Why should we pay much attention to information directed towards drivers in general if [we believe] we are safer and more skillful than they are?" But if drivers are only apparently overconfident, they may well pay attention. After all, a driver who thinks that she is above average with probability 0.55, also thinks she is below average with probability 0.45. If agents are truly overconfident, there may be excessive rates of entry, leading to negative profits, there may be overtrading (Barber and Odean, 2000; Glaser and Weber, 2007), and there may be too many lawsuits, labor strikes, and wars (Johnson and Fowler, 2011). In contrast, apparent overconfidence does not carry these negative implications, as agents who are only apparently overconfident are processing the available information correctly.[21]

There are several reasons for making the distinction between overconfidence and a preference for betting on oneself. To name a few, if people are overconfident, supplying them with better information about themselves may change their behaviour; if, instead, their choices reflect a desire to exert control, better information is irrelevant. If people enjoy betting on themselves there are few adverse welfare implications, at least in situations where no one else is implicated. An overconfident person might quit her current job because she overestimates her prospects in another job, but there is no particular reason for a preference for exerting control to push in that direction.

7 Conclusion

Experiments on overconfidence involve trade-offs. Walton (1999) asked truckers if they considered themselves to be safer drivers than the average trucker. This methodology has numerous flaws, including that the truckers were not incentivized, it is not clear that they all agreed on the definition of a safe driver, and it is not clear what a trucker's answer meant. The methodology has virtues, as well. The issue of safety is of substantial importance to truckers and they can be expected to have given it some thought even before the experiment. The comparative question "Are you better than most others?" is easy to understand. Even the more precise question, "Is it likely that you are better than most others?" is fairly natural.

[21] Benoît and Dubra (2007) show that in the experimental setup of Camerer and Lovallo (1999), apparent overconfidence can lead to increased entry without causing negative expected profits. This theoretical prediction is consistent with Camerer and Lovallo's experimental findings.

Unfortunately, the theory of B&D shows that this type of question, by itself, is of limited use. Arguably, the next most natural type of question is of the form "Are you likely to be better than $\frac{2}{3}$ of the others?" or "Are you quite confident, say with probability at least 0.75, that you are better than most others?" Our experimental results suggest that questions of this type, even incentivized, are unlikely to find overconfidence that passes Theorem 1's hurdle. However, Theorem 1 tests are quite conservative and it is entirely possible for them to fail to detect true overconfidence that is too small to surpass their demanding criteria.

Stronger tests of overconfidence ask subjects to provide specific probability estimates. However, questions such as these place a larger cognitive burden on subjects. It is easy to imagine that a subject who supplies a figure of, say, 0.7 might, if pushed, concede that 0.62 was an equally plausible estimate. Moreover, incentive compatible methods for eliciting specific probability measures tend to be unintuitive and often do not penalize participants very strongly for incorrect estimates. These disadvantages need to be weighed against the tighter tests that more complete information enables.

Another route uses actual placements to provide additional information on which to base studies. This route is promising but has its own pitfalls. In particular, it may involve tests that are of less importance to subjects than safety is to truckers and which introduce their own potential sources of error into the experiment.

Our two experiments, on easy quizzes, find overplacement. More precisely, we find apparently overconfident data that cannot be accounted for by a rational population of expected utility maximizers who care only about money and who have a good understanding of the nature of the quizzes they took.

Although our results point to overconfidence, our belief is that the jury is still out on the big question of how common overplacement actually is and how substantial the effect is. While there is a large body of experiments establishing the better-than-average effect on easy tasks, the body of experiments that employ a proper test of overplacement is quite small. The results in this literature are more mixed, with some experiments showing overplacement and others finding none. In any case, it is important to realize that the degree of (true) overconfidence may not be well measured by the fraction of people who rank themselves as above average.

8 Appendix A: Test items from the two tests

A complete list of test items can be found at learnmoore.org/mooredata/OJD/. Four randomly selected items are:

1) Susie has a cake that she splits into six pieces to share with all her friends. If each person with a piece of cake then splits their piece in half to give to another friend, how many pieces of cake are there in the end? 12

2) Fall is to Summer as Monday is to _____? Sunday

3) If two typists can type two pages in five minutes, how many typists will it take to type twenty pages in ten minutes? 10

4) There are three 600 ml water bottles. Two are full, the third is 2/3rds full. How much water is there total? 1600ml

9 Appendix B: Proofs

The following proposition shows that if the data passes a test based on Theorem 3, it also passes one based on Theorem 1.

Proposition 1 *Suppose that in a population of n individuals, r_i, $i = 1, ..., n$, is the probability with which individual i believes his type is in the top y, and that, in that same population, a fraction x of the population believe that there is a probability at least q that their types are in the top $y < q$ of the population. If $\frac{1}{n}\sum_{i=1}^{n} r_i = y$ then $qx \leq y$*

Proof. Let $Z = \{i \in \{1, ..., n\} \mid i \text{ believes there is a probability at least } q \text{ he is in top } y\}$. Then,

$$y = \frac{1}{n}\sum_{i=1}^{n} r_i = \frac{1}{n}\sum_{i \in Z} r_i + \frac{1}{n}\sum_{i \notin Z} r_i \geq \frac{1}{n}\sum_{i \in Z} r_i \geq \frac{1}{n}\sum_{i \in Z} q = qx,$$

as was to be shown. ■

Claim 1 *In experiment II, a person who overstates his probability of being in the top half by an additional* 26%*, has an expected loss of* \$0.33

Proof. The mechanism in Experiment II can be summarized as follows: the individual says an even number n between 0 and 100; the computer selects a number $x \in [0, 100] \cap \mathbf{N}$. If $x > n$, he wins \$10 with probability $\frac{x-1}{100}$ (individual draws a bingo ball, from a cage with 100 balls, and if it is lower than x you win \$10); if $x \leq n$, he wins \$10 if his score is in the top half.

The value of reporting n when the belief is b is

$$v(b, n) = \sum_{z=0}^{z=\frac{n}{2}} \frac{1}{51}\frac{b}{100}10 + \sum_{z=\frac{n}{2}+1}^{z=50} \frac{1}{51}\frac{2z-1}{100}10 = \frac{1}{1020}bn - \frac{1}{2040}n^2 + \frac{1}{510}b + \frac{250}{51}$$

It is easy to check that the mechanism elicits the closest even number to the individual's belief. Suppose the individual considers overstating his belief b by $M = 26$. Then, $v(b, b) - v(b, b + M) = \frac{M^2}{2040} = 0.33$ cents. ■

10 Appendix C: Statistical Tests

In this appendix, we develop our statistical tests for Theorem 2. The data for each test is a 5-tuple $(n, q, y, x, \widetilde{x})$, where n is the number of subjects, q the confidence with which they place in the top y of the population, x is the fraction who believe they are in the top y with probability at least q, and $\widetilde{x}$ is the fraction who have those beliefs and who actually place in the top y. We restrict ourselves to the case $xq > \widetilde{x}$; when $xq < \widetilde{x}$ the data passes a Theorem 2 test and it will easily pass a statistical test as well. When $xq > \widetilde{x}$, the data fails a Theorem 2 test and we would like to know the probability that it might, nevertheless, be a sample that was drawn from a larger population whose beliefs were generated by a rationalizing model.

Since n people can divide into at most n groups, it is without loss of generality to only consider rationalizing models with n types. It is also without loss of generality to make the number of signals the same as the number of types. Thus, we can describe a rationalizing model by a probability matrix $[p_{ij}]_{i,j=1,\ldots,n}$ where p_{ij} is the probability of drawing a type θ_j agent who sees a signal s_i.

Given a rationalizing model from which a sample of n individuals is drawn, we denote by $P\left(n, q, y, x, \tilde{z} \leq \widetilde{x}\right)$ the probability that, in that sample, a fraction x believe they are in the top y with probability at least q, and a fraction $\widetilde{x}$ or smaller have those beliefs and are actually in the top y. Recall that when the data $(n, q, y, x, \widetilde{x})$ fails a Theorem 2 test, it is because $\tilde{x}$ is too small. Thus, $P\left(n, q, y, x, \tilde{z} \leq \widetilde{x}\right)$ is the probability that a sample as “bad or worse” than $(n, q, y, x, \widetilde{x})$ could have come from within a rationalizing model. Our null hypothesis is that subjects are rational, so we want to choose the rationalizing model that maximizes $P\left(n, q, y, x, \tilde{z} \leq \widetilde{x}\right)$. We now proceed to characterize that model.

Let S_H denote the set of signals that induce a subject to believe that he is in the top y with probability at least q. We have,

$$s_h \in S_H \Leftrightarrow P\left(\theta \geq \theta_{n+1-y} \mid s_h\right) = \frac{\sum_{j=n+1-y}^{n} p_{hj}}{\sum_{j=1}^{n} p_{hj}} \geq q. \tag{1}$$

Also,

$$P\left(S_H\right) = \textstyle\sum_{s_h \in S_H} P\left(s_h\right) = \sum_{h: s_h \in S_H} \sum_{j=1}^{n} p_{hj} \tag{2}$$

The probability of a person believing he is in the top y and placing there is,

$$q_y = P\left(S_H \ \& \ \theta_j \geq \theta_{n+1-y}\right) = \textstyle\sum_{s_h \in S_H} P\left(s_h \ \& \ \theta_j \geq \theta_{n+1-y}\right) = \sum_{h: s_h \in S_H} \sum_{j=n+1-y}^{n} p_{hj} \tag{3}$$

Moreover, (1), (2), and (3) imply:

$$q_y = \textstyle\sum_{h: s_h \in S_H} \sum_{j=n+1-y}^{n} p_{hj} \geq \sum_{h: s_h \in S_H} q \sum_{j=1}^{n} p_{hj} = qP\left(S_H\right) \tag{4}$$

The probability of a person believing he is in the top y and not placing there is,

$$q_f = P\left(s_h \in S_H \ \& \ \theta_j < \theta_{n+1-y}\right) = \textstyle\sum_{s_h \in S_H} P\left(s_h \ \& \ \theta_j < \theta_{n+1-y}\right) = \sum_{h: s_h \in S_H} \sum_{j=1}^{n-y} p_{hj}. \tag{5}$$

We now characterize $P\left(n,q,y,x,\tilde{z}\leq\widetilde{x}\right)$ for any given rationalizing model.

Lemma 1 *Take a rationalizing model* $\left[p_{ij}\right]_{i,j=1,\ldots,n}$*, with* q_y *and* q_f *defined as in (3) and (5). We have,*

$$P\left(n,q,y,x,\tilde{z}\leq\widetilde{x}\right)=\textstyle\sum_{k=nx-n\widetilde{x}}^{nx}\displaystyle\frac{nx!}{k!\left(nx-k\right)!}\left(\frac{q_f}{q_f+q_y}\right)^{k}\left(\frac{q_y}{q_f+q_y}\right)^{nx-k} \tag{6}$$

Proof. First note that for any $\tilde{z}\leq nx$, the probability that in a sample of n the model generates the data $\left(n,q,y,x,\frac{\tilde{z}}{n}\right)$ is $\frac{n!}{\tilde{z}!(nx-\tilde{z})!(n-x)!}q_f^{nx-\tilde{z}}q_y^{\tilde{z}}\left(1-q_y-q_f\right)^{n-nx}$. Summing from 0 to nx, we can write the probability of drawing a sample in which nx individuals believe they are above y as

$$\begin{aligned}
P\left(x\right) &= \sum_{\tilde{z}=0}^{\tilde{z}=nx}\frac{n!}{(nx-\tilde{z})!\tilde{z}!\left(n-nx\right)!}q_f^{nx-\tilde{z}}q_y^{\tilde{z}}\left(1-q_y-q_f\right)^{n-nx}\\
&= \frac{n!}{(nx)!\left(n-nx\right)!}\left(1-q_y-q_f\right)^{n-nx}\left(\sum_{\tilde{z}=0}^{\tilde{z}=nx}\frac{(nx)!}{(x-\tilde{z})!\tilde{z}!}q_f^{nx-\tilde{z}}q_y^{\tilde{z}}\right)\\
&= \frac{n!}{(nx)!\left(n-nx\right)!}\left(q_f+q_y\right)^{nx}\left(1-q_y-q_f\right)^{n-nx}\sum_{\tilde{z}=0}^{\tilde{z}=nx}\frac{(nx)!}{(nx-\tilde{z})!\tilde{z}!}\left(\frac{q_f}{q_f+q_y}\right)^{nx-\tilde{z}}\left(\frac{q_y}{q_f+q_y}\right)^{\tilde{z}}\\
&= \frac{n!}{(nx)!\left(n-nx\right)!}\left(q_f+q_y\right)^{nx}\left(1-q_y-q_f\right)^{n-nx}
\end{aligned}$$

The last equality follows because the last sum above is the sum of all terms of a binomial distribution $B\left(nx,\frac{q_y}{q_n+q_y}\right)$.

Therefore, $P=P\left(n,q,y,x,\tilde{z}\leq\widetilde{x}\right)$ is

$$\begin{aligned}
P &= \frac{\sum\limits_{k=nx-n\widetilde{x}}^{k=nx}\frac{n!}{k!(nx-k)!(n-nx)!}q_f^{k}q_y^{nx-k}\left(1-q_y-q_f\right)^{n-nx}}{\frac{n!}{(nx)!(n-nx)!}\left(q_f+q_y\right)^{nx}\left(1-q_y-q_f\right)^{n-nx}}=\frac{\sum\limits_{k=nx-n\widetilde{x}}^{k=nx}\frac{n!}{k!(nx-k)!(n-nx)!}q_f^{k}q_y^{nx-k}}{\frac{n!}{(nx)!(n-nx)!}\left(q_f+q_y\right)^{nx}} \qquad (7)\\
&= \frac{\sum_{k=nx-n\widetilde{x}}^{nx}\frac{nx!}{k!(nx-k)!}q_f^{k}q_y^{nx-k}}{\left(q_f+q_y\right)^{nx}}=\textstyle\sum_{k=nx-n\widetilde{x}}^{nx}\displaystyle\frac{(nx)!}{k!\left(nx-k\right)!}\left(\frac{q_f}{q_f+q_y}\right)^{k}\left(\frac{q_y}{q_f+q_y}\right)^{nx-k}
\end{aligned}$$

as was to be shown. ■

The next step is to find the rationalizing model that maximizes $P\left(n,q,y,x,\tilde{z}\leq\widetilde{x}\right)$. Given $\left(n,q,y,x,\widetilde{x}\right)$, this is done by by choosing p_{ij}'s so that the resultant q_f and q_y maximize the rhs of (6). Letting $a=\frac{q_y}{q_f+q_y}$, we can recast this as choosing a to maximize

$$\textstyle\sum_{k=nx-n\widetilde{x}}^{nx}\displaystyle\frac{nx!}{k!\left(nx-k\right)!}\left(1-a\right)^{k}a^{nx-k} \tag{8}$$

Note that the derivative of each term in (8) with respect to a is directly proportional to $-k\left(1-a\right)^{k-1}a^{nx-k}+\left(nx-k\right)\left(1-a\right)^{k}a^{nx-k-1}$ and

$$\begin{aligned}
-k\left(1-a\right)^{k-1}a^{nx-k}+\left(nx-k\right)\left(1-a\right)^{k}a^{nx-k-1} &< 0\Leftrightarrow\\
\frac{nx-k}{k} &< \frac{a}{1-a} \qquad (9)
\end{aligned}$$

From equation (4),

$$a=\frac{q_y}{q_f+q_y}=\frac{q_y}{p\left(S_H\right)}\geq q.$$

Since $qx>\widetilde{x}$, we have that for $k\geq nx-n\widetilde{x}$,

$$\frac{a}{1-a}\geq\frac{q}{1-q}>\frac{\frac{\widetilde{x}}{x}}{1-\frac{\widetilde{x}}{x}}=\frac{\widetilde{x}}{x-\widetilde{x}}\geq\frac{nx-k}{k},$$

so that (9) holds. Hence, (8) is decreasing in a. To maximize (6), we set a to its lowest possible value; that is, we set $a=q$. The probability of a sample as overconfident as our data or more then becomes

$$P\left(n,q,y,x,\tilde{z}\leq\widetilde{x}\right)=\textstyle\sum_{k=nx-n\widetilde{x}}^{nx}\displaystyle\frac{nx!}{k!\left(nx-k\right)!}\left(1-q\right)^k q^{nx-k} \tag{10}$$

Equation (10) provides the statistical test based on Theorem 2, and is the number reported in the last columns of Tables 3 and 5.

Table 5 below lists the data from Experiment II, organized to perform tests based on theorems 1 and 2. Reading across, for instance, the third row, the first entry indicates that people are placing themselves in the top 50%, the second entry indicates a probability of at least 60% of placing there, the third entry indicates that 72% of the subjects have stated a probability of at least 60% of placing there, the fourth entry multiplies together the second and third entry, the fifth entry indicates that 43% of the subjects have stated a probability of at least 60% of placing in the top half and have placed in the top half and the sixth entry indicates $P\left(n,q,y,x,\tilde{z}\leq\widetilde{x}\right)$, maximized over rationalizing models.

Table 5. Tests for Experiment II.

Based on →		Theorem 1			Theorem 2		
y	q	x	qx	Pass if $qx < y$	$\tilde{x}$	Likelihood using q%	Likelihood using $q - 1$%
0.5	0.5	0.90	0.45	Pass	0.47	$> 50\%$	$> 50\%$
0.5	0.58	0.73	0.42	Pass	0.45	$> 50\%$	$> 50\%$
0.5	0.60	0.72	0.43	Pass	0.43	$> 50\%$	$> 50\%$
0.5	0.66	0.61	0.40	Pass	0.36	24%	29%
0.5	0.68	0.59	0.40	Pass	0.35	14%	17%
0.5	0.70	0.55	0.39	Pass	0.32	8%	10%
0.5	0.72	0.43	0.30	Pass	0.24	4%	5%
0.5	0.74	0.42	0.31	Pass	0.24	4%	5%
0.5	0.76	0.40	0.31	Pass	0.23	2%	2%
0.5	0.78	0.36	0.29	Pass	0.22	2%	3%
0.5	0.80	0.35	0.28	Pass	0.20	$< 1\%$	1%
0.5	0.84	0.23	0.19	Pass	0.15	4%	5%
0.5	0.86	0.20	0.17	Pass	0.15	15%	18%
0.5	0.88	0.19	0.17	Pass	0.13	8%	10%
0.5	0.90	0.18	0.16	Pass	0.12	3%	5%
0.5	0.92	0.08	0.07	Pass	0.05	8%	10%
0.5	0.94	0.07	0.06	Pass	0.05	3%	4%
0.5	0.96	0.05	0.05	Pass	0.04	15%	19%
0.5	0.98	0.04	0.04	Pass	0.04	$> 50\%$	$> 50\%$
0.5	1	0.03	0.03	Pass	0.03	$> 50\%$	$> 50\%$

Notes: x is the percentage with belief greater than q that they will place in the top half; $\tilde{x}$ is the fraction of x who actually place in the top half. The last two columns represent the probability that data comes from a rational model. They are $P\left(w \leq n\widetilde{x}\right)$, where w is a binomial $B\left(nx, q\right)$ (or $q - 1$).

References

Ariely, D., U. Gneezy, G. Loewenstein and N. Mazar, (2005), "Large stakes and big mistakes," available at SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=774986

Barber, B. M., & Odean, T. (2000), "Trading is hazardous to your wealth: The common stock investment performance of individual investors," *Journal of Finance*, **55(2)**, 773–806.

Barber, B. and T. Odean (2001), "Boys Will Be Boys: Gender, Overconfidence, And Common Stock Investment," *Quarterly Journal of Economics*, 116(1), 261-92.

Baumeister, R. F., Campbell, J. D., Krueger, J. I., & Vohs, K. D. (2003), "Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles?," *Psychological Science in the Public Interest*, **4**, 1–44.

Beilock, S. L. and T.H. Carr (2001), "On the fragility of skilled performance: What governs choking under pressure?" *Journal of Experimental Psychology: General*, **130(4)**, 701-25.

Bem, D.J. (1967), "Self-perception theory: An alternative interpretation of cognitive dissonance phenomena," *Psychological Review*, **74(3)**, 183-200.

Bénabou, R. and J. Tirole (2002), "Self Confidence and Personal Motivation," *Quarterly Journal of Economics*, **117(3)**, 871-915.

Benoît, Jean-Pierre and Dubra, Juan (2007): Overconfidence?, Munich Personal RePEc Archive, paper 5505.

Benoît, J-P. and J. Dubra (2011), "Apparent Overconfidence" *Econometrica* **79(5)**, 1591-1625.

Bernardo, A. and I. Welch (2001), "On the Evolution of Overconfidence and Entrepreneurs," *Journal of Economics & Management Strategy*, **10(3)**, 301-330.

Brocas, I. and J. Carrillo (2007), "Systematic errors in decision-making," mimeo.

Burks, S. J. Carpenter, L. Goette and A. Rustichini (2013), "Overconfidence and Social Signalling," *Review of Economic Studies*, **80(3)**, 949-83.

Camerer, C. and Lovallo, D. (1999). Overconfidence and excess entry: an experimental approach', *American Economic Review*, **89(1)**, pp. 306–18.

Clark, J. and L. Friesen (2009), "Overconfidence in Forecasts of Own Performance: An Experimental Study," *Economic Journal*, **119(1)**, 229-51.

Chuang, W. and B. Lee, (2006), "An empirical evaluation of the overconfidence hypothesis," *Journal of Banking & Finance*, 30(9), 2489-515.

Daniel, K., D. Hirshleifer and A. Subrahmanyam (2001), "Overconfidence, Arbitrage, and Equilibrium Asset Pricing," *Journal of Finance*, **56(3)**, 921-65.

Dohmen, T. J. (2005), "Do professionals choke under pressure?" Unpublished manuscript.

Dunning, D. (2005). Self-insight: Roadblocks and detours on the path to knowing thyself. New York: Psychology Press.

Eil, D. and J. Rao (2011), "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself," *American Economic J: Microeconomics,* **3**, 114-38.

Fang, H. and G. Moscarini, (2005) "Morale Hazard," *J. of Monetary Economics*, **52(4)**, 749-777.

Festinger, L. (1954) "A Theory of Social Comparison Processes," *Human Relations*, **7(2)**, 117-40.

Garcia, D., F. Sangiorgi and B. Urosevic, (2007), "Overconfidence and Market Efficiency with Heterogeneous Agents," *Journal Economic Theory*, **30(2)**, 313-36.

Glaser, M., & Weber, M. (2007), "Overconfidence and trading volume," *Geneva Risk and Insurance Review*, 32, 1–36.

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

Goodie, A. and D. Young (2007), "The skill element in decision making under uncertainty: Control or competence?," *Judgment and Decision Making*, **2(3)**, pp. 189-203.

Grether, D. M. (1981) "Financial Incentive Effects and Individual Decision Making," Social Sciences working paper 401, Cal. Tech.

Harris, A. J. L. and U. Hahn (2011), "Unrealistic optimism about future life events: A cautionary note." *Psychological Review*, **118**, 135-54.

Heath, C. and A. Tversky, (1991) "Preference and Belief: Ambiguity and Competence in Choice under Uncertainty," *Journal of Risk and Uncertainty*, **4**, 5-28.

Hoelzl, E. and A. Rustichini, (2005), "Overconfident: do you put your money on it?" the *Economic Journal*, **115**, pp. 305-18.

Johnson, D. D. P., and Fowler, J. H. (2011), "The evolution of overconfidence," *Nature*, 477(7364), 317–320.

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

Karni, E. (2009), "A Mechanism for Eliciting Probabilities," *Econometrica*, **77(2)**, 603–6.

Klar, Y., and Giladi, E. E. (1999), "Are most people happier than their peers, or are they just happy?" *Personality and Social Psychology Bulletin*, 25(5), 585–594.

Kőszegi, B., (2006), "Ego Utility, Overconfidence, and Task Choice," *Journal of the European Economic Association*, **4(4)**, 673-707.

Kruger, J., & D. Dunning (1999), "Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments," *Journal of Personality and Social Psychology*, **77**, 1121–34.

Kunda, Z. (1990). The case for motivated reasoning. Psychological Bulletin, 108(3), 480-498.

Kyle, A. and F.A. Wang, (1997), "Speculation Duopoly with Agreement to Disagree: Can Overconfidence Survive the Market Test?" *Journal of Finance*, **52(5)**, 2073-90.

Larrick, R. P., K. A. Burson and J.B. Soll, (2007), "Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not)," *Organizational Behavior and Human Decision Processes*, **102**, 76–94.

Logg, J. M., Moore, D. A., & Haran, U. (2013). Motivation and overconfidence. Unpublished Manuscript.

Malmendier, U. and G. Tate (2005), "CEO Overconfidence and Corporate Investment," *Journal of Finance*, **60(6)**, 2661-700.

Markman, A. B., W.T. Maddox, (2006), "Choking and excelling under pressure," *Psychological Science*, **17(11)**, 944-48.

Menkhoff, L., U. Schmidt and T. Brozynski, (2006) "The impact of experience on risk taking, overconfidence, and herding of fund managers: Complementary survey evidence," *European Economic Review*, **50(7)**, 1753-66

Merkle, C. and M. Weber (2011), "True Overconfidence: The Inability of Rational Information Processing to Account for Apparent Overconfidence," *Organizational Behavior and Human Decision Processes*, **116(2)**, 262-71.

Moore, D. A. (2007), "Not so above average after all: When people believe they are worse than average and its implications for theories of bias in social comparison," *Organizational Behavior and Human Decision Processes*, **102(1)**, 42–58.

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

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

Peng, L. and W. Xiong, (2006), "Investor attention, overconfidence and category learning," *Journal of Financial Economics*, **80(3)**, 563-602.

Plott, C. and K. Zeiler (2005), "The Willingness to Pay-Willingness to Accept Gap, the Endowment Effect, Subject Misconceptions, and Experimental Procedures for Eliciting Valuations," *American Economic Review,* **95(3)**, pp. 530-45.

Pulford, B. D., & Colman, A. M. (1997). Overconfidence: Feedback and item difficulty effects. Personality and Individual Differences, 23(1), 125–133.

Svenson, O., (1981), "Are we all less risky and more skillful than our fellow drivers?" *Acta Psychologica*, **94**, pp 143-148.

van den Steen, E. (2004), "Rational overoptimism (and other biases)," *American Economic Review*, **94(4)**, 1141–51.

Walton, D., (1999), "Examining the self-enhancement bias: professional truck drivers' perceptions of speed, safety, skill and consideration," *Transportation Research Part F,* 91-113.

Wang, A. (2001), "Overconfidence, Investor Sentiment, and Evolution," *Journal of Financial Intermediation*, **10(2)**, 138-70.

Weinstein, N. (1980), "Unrealistic Optimism about Future Life Events," *Journal of Personality and Social Psychology*, **39(5)**, 806-20.

Windschitl, P., Rose, J., Stalkfleet, M., & Smith, A. (2008), "Are people excessive or judicious in their egocentrism? A modeling approach to understanding bias and accuracy in people's optimism within competitive contexts," *J. of Personality and Social Psychology*, **95(2)**.

Zábojník, J. (2004), "A Model of Rational Bias in Self-Assessments," *Economic Theory,* **23(2)**, 259–82.