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

Benoît, Jean-Pierre and Dubra, Juan and Moore, Don

Universidad de Montevideo

11 October 2009

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 experiements of the claim that people are overconfident. We develop new tests of overplacement which are based on a formal Bayesian model. 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 with a good understanding of the nature of the quizzes they took.

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, at least on easy tasks.[1] For

*Authors are listed alphabetically. 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.

†email: dubraj@um.edu.uy

[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 financial markets, Kőszegi (2006) who studies how overconfidence affects how 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.

economists, the issue of overconfidence is of paramount importance as it affects the equilibrium outcomes in almost every market. Although the term "overconfidence" has been used rather broadly, Larrick, Burson, and Soll (2007) and Moore and Healy (2008) point out that, in fact, three distinct varieties of overconfidence have been examined in the literature: (1) people having excessive confidence in the their estimates, or *overprecision*, (2) people *overestimating* their abilities, and (3) people *overplacing* themselves relative to others. In this paper, we focus on the third type of overconfidence, overplacement.

For the most part, researchers have not directly observed overplacement but have, instead, inferred it from the tendency of a majority of people to claim to be superior to the median person – the so-called *better-than-average effect*. The better-than-average-effect has been noted for a wide range of easy skills, from driving, 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 recently questioned its significance. They show that better-than-average data in and of itself merely gives the *appearance* that (some) people must be overplacing themselves, but does not indicate *true* overplacement, which carries with it the implication that people have made some kind of error in their self-placements.[3] Because of this reason, the vast majority of the existing experimental literature on the better-than-average effect cannot actually claim to have found overplacement. Moreover, most of the experiments by their very design do not even have the potential of showing overplacement. In this paper, we report on two experiments which provide a proper test of overplacement.

The most common type of experiment in this field involves asking subjects, either explicitly or implicitly, how they rank compared to others. For instance, Svenson (1981), in perhaps the most cited study, asks subjects to estimate how their driving compares to the others by placing themselves into one of ten successive ten percent intervals; 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. There are at least three criticisms that can be made of this type of experiment, though not every criticism applies to every experiment:[4]

1. Participants often have no material incentive to answer the question accurately and internal motivations to answer accurately are likely to compete with other motivations,

[2] While early research pointed towards a universal better-than-average effect, 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).

[4] These are criticisms of the experiments as tests of overconfidence. Many of these experiments have other purposes as well, that may not be subject to these criticisms. For instance, Hoelzl and Rustichini (2005) are also interested in understanding the better-than-average effect per se and the extent to which it survives various manipulations.

such as appearing competent, self-confident, or modest.

2. Subjects may be uncertain of their skill levels, making the meaning of their answers unclear.

3. The underlying theory that leads to the conclusion of overconfidence has not been carefully delineated, and the implicit theory is often erroneous.

The first criticism is quite familiar, so let us turn to the next two. Consider a subject who is asked to rank himself on IQ, given that the median IQ is 100. If he has not actually taken an IQ test then he must guess at his IQ. Suppose that he believes that his IQ is 80 with probability 0.45, 110 with probability 0.45, and 115 with probability 0.1. How should he rank himself? He could reasonably respond that be believes himself to be of above average intelligence, given that there is over a 50% chance that his IQ is above average. On the other hand, he could just as reasonably respond that he is of below average intelligence, given that his mean IQ is only 97. Thus, the subject's answer to the question gives no clear indication of its meaning. By the same token, a statement like "I believe I have a higher IQ than the average person" gives no indication of the degree of confidence with which it was uttered.[5] As to the third criticism, suppose that 80% of subjects rank themselves above the median. Experimenters have simply asserted that this is evidence of overconfidence, without a careful articulation of why this is so. The following example, which is similar to one in B&D, illustrates the flaw in this approach.

Consider a large population with three types of drivers, low skilled, medium skilled, and high skilled, and suppose that the probabilities of any one of them causing an accident in any single period are $p_L = \frac{4}{5}, p_M = \frac{2}{5}$, and $p_H = 0$. In period 0, nature chooses a skill level for each person with equal probability. Initially, no driver knows his or her own skill level, and so each person (rationally) evaluates himself as no better or worse than average. In period 1, everyone drives and learns something about his skill, based upon whether or not he has caused an accident. Each person is then asked how his driving skill compares to the rest of the population. How does a driver who has not caused an accident reply?

[5] Note, however, that if, as a matter of fact, subjects are very sure of their types then these issues become moot – the various meanings that subjects could have for their answers converge. In addition to testing for overconfidence we test the hypothesis that subjects are not very sure of their types.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 types. Several strands of the psychology literature also stress that people are uncertain of their types, including Festinger's (1954) influential social comparison theory, Bem's (1967) self-perception theory, and Amabile (1983).

Using Bayes' rule, he evaluates his own skill level as follows:

$$\begin{aligned} p\,(\text{Low skill} \mid \text{No accident}) &= \frac{\frac{1}{3}\frac{1}{5}}{\frac{1}{3}+\frac{1}{3}\frac{3}{5}+\frac{1}{3}\frac{1}{5}} = \frac{1}{9} \\ p\,(\text{Medium skill} \mid \text{No accident}) &= \frac{\frac{1}{3}\frac{3}{5}}{\frac{1}{3}+\frac{1}{3}\frac{3}{5}+\frac{1}{3}\frac{1}{5}} = \frac{1}{3} \\ p\,(\text{High skill} \mid \text{No accident}) &= \frac{\frac{1}{3}}{\frac{1}{3}+\frac{1}{3}\frac{3}{5}+\frac{1}{3}\frac{1}{5}} = \frac{5}{9} \end{aligned}$$

Such a driver thinks there is over a $\frac{1}{2}$ chance (in fact, $\frac{5}{9}$) that his skill level is in the top third of all drivers. His mean probability of an accident is $\frac{5}{9}0+\frac{1}{3}\frac{2}{5}+\frac{1}{9}\frac{4}{5}=\frac{2}{9}$, which is better than for $\frac{2}{3}$ of the drivers, and better than the population mean. Furthermore, his beliefs about himself strictly first order stochastically dominate the population distribution. Any way he looks at it, a driver who has not had an accident should evaluate himself as better than average. Since $\frac{3}{5}$ of drivers have not had an accident, $\frac{3}{5}$ rationally rank themselves as better than average.

As this example shows, the fact that 60% of drivers rank themselves above the median does not indicate erroneous self-evaluations. In fact, Theorem 1 below shows that any fraction less than one of people could rationally believe that they have over a 50% chance of ranking in the top half of the population, without any overplacement being implied. Therefore, any experiment designed just to show that more than half the population rank themselves as likely to be better than then median cannot possibly show overplacement. Experiments with more detailed information on how subjects place themselves in percentiles have the potential to show overplacement, but even these must be carefully interpreted.

We conduct two experiments that enable us to perform a variety of tests with the potential to show if people are not making rational assessments of their abilities. Both our experiments find overconfidence, though not all the tests we run reveal this overconfidence. In Section 5, we discuss some possible limitations of these findings. Subject to these caveats, our results join those of Merkle and Weber (2011) and Burks et. al. (2011) who also conduct proper tests and find overplacement (and are subject to similar caveats). Two experiments that conduct proper tests of overplacement, but do not find such a bias, are Clark and Friesen (2008) and Moore and Healy (2008).[6]

[6]Merkle and Weber (2011) and Burks et al. (2011) take explicit account of the critique of B&D. Clark and Friesen (2008) and Moore and Healy (2008) preceded B&D, but their implicit theory of overplacement is correct (and, in fact, corresponds to Theorem 3 of B&D).

2 Background

When should we say that a person is *overconfident*? An immediate proposal is that an overconfident person is not as skillful as she thinks she is. However, making such a determination may be problematic, as many skills are not easily measured. For instance, consider a person who asserts "I am a very good driver". Even supposing that we can make the notion of "very good" precise and that we can agree on what constitutes a very good driver, how are we to determine if the statement is true? Giving the person a driving test may not be practical. Moreover, the skills measured in such a test may not match up very well with the day-to-day skills reflected in the driver's self-assessment.

Researchers have circumvented these problems by considering entire populations at once and asking subjects how their skills compare to each other. Beyond circumvention, there are at least two reasons to be interested in this overplacement. Firstly, in many domains people may well have a better idea of their relative placements than their absolute placements. Thus, we might expect students to have a better idea of their math abilities relative to their classmates, than of their absolute abilities. Secondly, in many areas of interest, relative ability is of primary importance. For instance, in many jobs success depends primarily on a person's abilities relative to his or her peers.

The basic idea behind the relative population approach is that, since only 50% of people can be in the top 50% in skill level, if more than half the people in a population claim to be in the top half – or make choices which reveal such a belief – they "must" be making an error. However, as the example in the introduction shows, this idea is flawed.

The implication in terming a population overconfident is that the members of the population have made some errors or have some inconsistencies in their self-evaluations.[7] Thus, B&D proposes that data be called (truly) overconfident only if it cannot be obtained from a population that derives its beliefs in a fully rational and consistent manner, as follows.

Let a **rationalizing model** be a four-tuple $\left(\Theta, p, S, \{f_\theta\}_{\theta\in\Theta}\right)$, where $\Theta \subseteq \mathbf{R}$ is a type space, p is a prior probability distribution over Θ, S is a 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 interpretation of the model is the following. 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

[7] These errors can be expected to lead to further errors, such as too many people attempting to become professional athletes.

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 a rationalizing model.

Since experimental data that is typically available does not reflect agents' complete beliefs, data that can be rationalized may still come from a population that is not acting in a fully rational and consistent manner. However, data that cannot be rationalized is definitely problematic. A proper test for overconfidence involves a search for data that cannot be rationalized. The nature of the test that can be conducted depends on the type of data that is available. Conversely, the type of data that an experimenter collects will depend upon the test that is to be conducted. In this paper, we report on four different tests. These tests are based on the three theorems below.

- We say that a person of type t is in the top y of a population if the fraction of people whose type is greater than or equal to t is at most y. Thus, in a population of 100 people at most 25 can be in the top $\frac{1}{4}$.

- Given a population of subjects, let $x(q, y)$ be the fraction who believe that there is a probability at least q that their types are in the top y of the group. If the subjects had no information pertaining to their individual abilities, then they all would believe that they had a probability y of ranking in the top y. Thus, for $q = y$ even a population in which $x(q, y) = 1$ would not demonstrate overconfidence, and a fortiori, neither would a population of subjects who were less confident that they were in the top y (i.e., $q < y$). Since we are interested in the possibility of overconfidence, we conduct experiments in which $q > y$ and state our theorems accordingly. Theorems for the case $q \leq y$ are essentially symmetric.

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. This data can be rationalized if and only if $qx \leq y$.*

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 $qx \leq \tilde{x}$.*

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 $\frac{1}{n}\sum_{i=1}^{n} r_i = y$.*

All theorems are proved in the appendix, though theorems 1 and 3 are essentially corollaries of theorems found in B&D.[8] Theorem 2 uses information about subjects' actual placement in addition to their beliefs. The first paper that we know of that uses actual placement data to conduct a proper test of overconfidence, albeit in a different way, is Burks et al. (2011).

Walton (1999) asks truckers whether or not they consider themselves to be safer drivers than the average truckers. However, the truckers are not incentivized and the meanings of their answers are unclear. The limitations of this study notwithstanding, it has its virtues. The question asked is of importance to the truckers and they can be expected to have given it some thought even prior to the study. Even the more precise questions, "Are you probably a safer driver than most?" or "Do you think that you are less likely to end up in an accident than most drivers" are fairly natural.

Unfortunately, as Theorem 1 shows, the questions are of little use for the study of overconfidence; more information is needed. A straightforward extension of the question is to narrow the range of placement to the top $y < \frac{1}{2}$, rather than the entire top half, or to increase the confidence in the placement to some $q > \frac{1}{2}$, or both. Beyond the question of overconfidence, part of our interest is the extent to which data like this, which is similar in spirit to better-than-average type data, can revel overplacement. Our findings are negative in this regard, suggesting that more information may be needed for a proper study, and we do conduct tests that use more information. However, there is often a trade-off between the amount of information collected, and the difficulty of the mechanism and the strength of the incentives, which may make the more detailed information less reliable. We return to this issue in Section 5.

Let us 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. 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 $(\tilde{x}, x, q)$ 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 new information. Similarly, if the data passes a test based on Theorem 3 it also passes one based on Theorem 1 (see the appendix), so that again the latter is made redundant. On the other hand, tests based on theorems 2 and 3 are independent of each other.

The above reasoning suggests that Theorem 1 may be of limited use. However, there

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

are several factors that make this theorem valuable. First, tests based on Theorem 1 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 the subject pool in their driving skill level. From Theorem 1, these drivers display overconfidence if we make the plausible, though by no means certain, assumption that their placement indicates an at least 50% belief that they are in the top 30%,[9] and we accept the validity of unincentivized responses, as many psychologists do. Second, at a practical level, there may be reasons to have more confidence in data elicited for tests based on Theorem 1 than in the more detailed data elicited for tests based on Theorem 2 or 3. Third, there are theoretical issues pertaining to the demands of rationalization that may favour tests based on Theorem 1. We discuss points two and three in Section 5.

3 Experiment I

From Theorem 1, we can infer overconfidence if a sufficient fraction of people (variable x in the theorem) believe sufficiently strongly (variable q) that they rank sufficiently high (variable y). From Theorem 2, we can infer overconfidence if too few people who rank themselves high actually place high. In our first experiment, we conduct tests of overplacement based on these two theorems. First, we test if more than 60% of the subjects believe that there is at least a 50% chance that their type is in the top 30%. Recall that Svenson found that over 80% of his American subjects placed themselves in the top 30%, but it was unclear what they meant by this placement. We also test if too many subjects feel that there is more than a 60% chance that they are better than the median. We choose 60% because we are independently interested in whether a relatively small increase in the chance of receiving a prize randomly – from 50% in a benchmark test to 60% here – makes many people change their choice behavior. Finally, we compare the beliefs of the subjects to their actual placements and check whether these beliefs are consistent with Theorem 2.

We were interested in the extent to which previous findings of apparent overplacement could be shown to be actual overplacement. Prior experimental work and the theory in B&D demonstrate that populations exhibit the better-than-average effect more markedly on easy tasks than difficult ones.[10] Accordingly, we gave our subjects an easy quiz.

Subjects are 134 individuals recruited through the web site of the Center for Behavioral

[9] Other possible interpretations include that subjects' responses reflect their mean beliefs about their abilities or their modal beliefs.

[10] The theory in Moore and Healy (2008) predicts that a test that is *easier than expected* should yield more overconfident looking data.

Decision Research at Carnegie Mellon University <http://cbdr.cmu.edu/experiments/>. We report the data for the 129 subjects who gave complete responses to the three choices with which they were presented; the results are unchanged when we analyze, for each question, all the answers we have for that question.

The experiment was advertised under the name "Test yourself" along with the following 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 wording of the recruitment instructions was chosen to be conductive to more "overconfident looking data" (Camerer and Lovallo (1999) find that excess entry into their game (their measure of overconfidence) is much larger when subjects volunteer to participate in the experiment knowing that payoffs will depend on skill).

Subjects had a mean age of 25 years (SD = 6.4) and 42 percent of them were male. All subjects saw a sample test. Then, they made a series of three choices between (1) bets on their test performance (skill) and (2) chance gambles of known probability. Subjects had to choose one of the two for each of the three pairs of bets. Finally, they took a 20-item quiz of math and logic puzzles. The three pairs of bets are listed below.

Skill Option

1. 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.

.

.

2. 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.

.

.

3. 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

.

.

Chance Option

1. 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. 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. 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.

Subjects were randomly assigned to experimental conditions that crossed two treatment variables: motivation and feedback.

The motivation manipulation varied what subjects were told about the test they were about to take. By introducing a manipulation of motivation we hoped to observe the effect of inducing a motive to be overconfident. Many theories of overconfidence assume that the belief that one is better than others is driven by the desire to actually be better than others (Benabou & Tirole, 2002; Kőszegi, 2006; Kunda, 1990). Therefore, people's propensity to overplace their performances relative to those of others ought to be greatest under those circumstances when they are most motivated to achieve (see Krizan & Windschitl, 2007). Those in the high motivation condition 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 IQ 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 & Deary, 2004)."

Those in the low motivation condition read: "In this experiment, you will be taking a test of math and logic puzzles."

Then subjects saw a set of sample test items. In order to constitute this set of sample items, we began with a larger set of 40 test 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.

Half of the subjects (those in the feedback condition) received a histogram showing how others had scored on the test they were about to take.

Next, subjects chose between skill and chance options for each of three bets. The order in which the three bets appeared was varied randomly, as was whether the chance or the skill option appeared first for each bet. 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.[11] The choices can be summarized as:

1. **Benchmark Treatment**: A 50% chance of a prize (as determined by a random draw), or to be awarded the prize if your score on the test places you in the top 50% of previous test takers.

2. **High Placement Treatment**: A 50% chance of a prize (as determined by a random draw), or to be awarded the prize if your score on the test places you in the top 30% of previous test takers.

3. **Strength Treatment**: A 60% chance of a prize (as determined by a random draw), or to be awarded the prize if your score on the test places you in the top 50% of previous test takers.

Then subjects took the twenty-item test under a ten-minute time limit. The two test sets appear in Appendix A. Subjects earned $.25 for each test question they answered correctly.

[11] The results we present are those of the first set of choices; those made before taking the test. This is the standard methodology for studying overplacement (see Moore and Healy (2008), Clark and Friesen (2008) and 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, and we do not present this data.

Then subjects chose between the skill and chance options for each of the three bets again. Subjects then answered a series of questions regarding what they thought their score would be, how they felt during the experiment, etc.

Finally, if a subject chose to bet on chance (rather than their test performance) for the one bet that counted, an experimenter had the subject draw from the relevant bag of poker chips to determine whether he or she won the $10 prize.

3.1 The data

There are 5 variables, none of which had any effect on the choice behavior of subjects (or their scores – except for the High Motivation treatment, which decreased scores, see below).

First, as expected, neither of the following three randomizations had any effect:

- The order of the presentation of the bets (123, 132, 213, etc).
- Whether the skill or random bet was presented first in each pair.
- Whether subjects saw sample M and took test S, or saw S and took M.

Second, we didn't have a prior belief of how the feedback manipulation would affect scores or choices between bets; it had no effect. Finally, and surprisingly to us, the Motivation manipulation had no effect either. Hence, we discuss only aggregate data, without discriminating by 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. In the Benchmark Treatment, 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%. In the High Placement Treatment, 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%. In the Strength Treatment, 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%.

As expected, in the Benchmark Treatment, the population displays apparent overplacement: 74% choose to be rewarded based upon their placement. Barring too many equally skilled subjects (and ignoring the possibility of errors), such a result is usually interpreted as 74% place themselves in the top half of test takers. However, this statement is imprecise, if not misleading. 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 are different and have different implications for rationality. In 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. Overplacement can be inferred only if the subjects' belief that they are in the top half is sufficiently more than 50% or if they believe they place sufficiently high within the top half.

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% random draw, 22% switch and choose a 60% random draw over placing in the top half.[12] 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.[13] In particular, it suggests that prior work on overconfidence cannot be justified by an untested presumption that people are certain, or nearly certain, of their types.

We turn now to the question of overplacement.[14]

Tests Based on Theorem 1.

From Theorem 1, the population exhibits overconfidence if more than 60% vote for the skill bet in the High Placement pair of bets, or if more than 83.3% vote for the skill bet in the Strength pair of bets. In fact, only 51.9% and not 60% choose the skill bet in the High Placement pair of bets, so that rationality cannot be rejected. More precisely, one can build a rational model in which a sample at least this overconfident looking arises with probability greater than 50% (i.e. a sample like this is very likely if the null is rationality; see Appendix C for details). If, on the other hand, the null hypothesis is that more than 60% believe that there is more than 50% chance of being in the top 30%, it can be rejected with 3%

[12] 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.

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

[14] Although it is not the focus of our study, we mention one intriguing finding. While the high/low motivation treatment does not affect the betting behaviour of our subjects, the subjects have significantly lower scores under the high motivation treatment. Those in the high motivation condition answered 16.6 questions correctly, whereas those in the low motivation condition answered an average of 18 questions 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 has been documented by other studies, including Ariely, Gneezy, Loewenstein, and Mazar (2005), Beilock and Carr (2001), Dohmen (2005), and Markman and Maddox (2006). In the present context, this finding is interesting in that it speaks to the potential adaptiveness (or lack thereof) of motivations to be confident.

significance (a t test with 128 degrees of freedom reveals that 51.9% is different from 60% at the 3% significance level). Also, only 64.3% and not 83.3% choose the skill bet in the Strength pair. Again, a sample as apparently overconfident as this, or more, has a likelihood greater than 50% in a rational model, and one can reject the null that more than 83.3% of the population believes there is a chance greater than 60% that they will score in the top half with a confidence greater than 99% (a t test with 128 degrees of freedom reveals that 64.3% is different from 83.3% at significance levels lower than 1%).

Although the results from these tests are consistent with no overconfidence, they do not rule out overconfidence since, perhaps inevitably, they only reflect a fraction of the subjects' beliefs. The fact that we have information on actual test performance allows us to conduct more stringent tests based on Theorem 2.

Tests Based on Theorem 2.

From Theorem 2:

1. At least 60% of the of subjects who bet on themselves in the strength treatment should have scores in the top half. In fact, only 54.9% of those who bet this way place in the top half.[15] However, while 54.9 is less than 60, the following statistical test reveals that in a rational model, there is a 16.7% chance that a sample as apparently overconfident as this, or more, will arise. Consider a rationalizing model, with 2 types, each with probability $\frac{1}{2}$, and 2 signals, and 64.3% observe the high signal (which has a posterior of 60% in the high type). In samples of 129 individuals, in which 64.3% observe the high signal, the chance that 54.9% or less of those individuals will score in the top half is 16.7% (so we can't reject rationality[16]).

2. At least half of the 52% of subjects who bet on themselves in the high placement treatment should have scores in the top 30%. In fact, only 32.8% (and not 50%) of those who bet this way do place in the top 30%.[17] The following statistical test

[15] Some care must be taken in determining the percentage who place in the top half, that is who place among the top 65 subjects. The median score is 18. There are 54 subjects who score more than 18 and 18 subjects who score exactly 18. Hence, 11 of the 18 who score 18 are randomly chosen to place in the top half. There are 14 individuals who both score 18 and claim to be in the top half, so that $14 * \frac{11}{18} \approx 9$ of them end up in the top half. Together with the 37 who claim to be in the top half and score above 18, we have that 46 out of the 83 (54.9%) who bet on their score actually placed in the top half.

[16] If we round the $14 * \frac{11}{18} = 8.5556$ of the previous footnote to 9, instead of the conservative 8, the probability of the sample increases to 22.9%.

[17] The top 30% of test takers is 39 and 21 score more than the cutoff for the 30^{th} percentile (cutoff is 19), while 33 score 19. Hence, 18 out of those 33 place in the top 30%. There are 17 individuals who score the cutoff score 19, and claim to be in the top half, so $17 * \frac{18}{33} \approx 9$ of them end up in the top 30%; together with the 13 who claim to be in the top 30% and score in that range, we have that 22 out of the 67 (32.8%) who bet on their score actually placed in the top 30%.

reveals that in the rational model that maximizes the chance of a sample as apparently overconfident as this one (or more), the likelihood of this much apparent overconfidence is less than 1%. Consider a two type- two signal model, in which the high type has probability 30%, and the high signal has a chance of 52% (which has a posterior of 50% in the high type). In samples of 129 individuals, in which 52% observe the high signal, the chance that 32.8% or less of those individuals will score in the top 30% is less than 1% (so we reject rationality).

Combining the results from tests based on theorems 1 and 2, the data passes three out of four tests. However, to be rational, the data must pass *all* tests. Thus, Experiment I rejects the hypothesis that subjects are behaving rationally, although the tests based just on Theorem 1 could not rule out rationality.

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 very similar in its overall design to Experiment I. It again involves Carnegie Mellon undergraduates, 74 this time, taking a quiz very similar to the previous ones. The crucial difference is that subjects were asked, in an incentive compatible manner, to indicate the likelihood they ascribed to placing in the top half. The elicitation mechanism used was the *probability matching rule* described by Karni (2009) and Grether (1981), as implemented by steps 3 and 6 below. The steps of the experiment were:

1. Participants took a five-item practice quiz. They had 2.5 minutes. A record was kept of their score.

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

3. 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.

4. 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.

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

6. The computer chose an even number uniformly from 2% to 100%. Participants who had indicated a number larger than that chosen by the computer, were rewarded according to whether their score was in the top half; those who had chosen a number equal or lower to that of the computer drew a bingo ball from a cage with even numbers from 2 to 100. If number on the ball was equal to lower than the number chosen by the computer, they won $10.

With the probability matching rule, it is optimal for expected utility maximizing subjects to report their true subjective probabilities when they can choose any number from the interval $[0, 100]$. There is a wrinkle in the experiment, however, as subjects and the randomizing device were both restricted to choosing even numbers. With this restriction, if a subject's subjective probability of success is not an even number, it is optimal for the subject to round up to the next highest even number, though this fact was not emphasized.

The reason for Step 4 is that we wanted to make sure that people who chose a very 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 4a over 4b.

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 almost up to 52% in the experiment. The actual average given was 67.2%, which is greater than 50% 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 which was gathered also enables us to conduct an additional nineteen tests based on Theorem 1 and sixteen based on Theorem 2. For instance, 35% of subjects indicated that they have a probability of at least 0.8 of ending in the top half. From Theorem 1, up to 62% of subjects could rationally make such an indication, so the data passes this test. (More precisely, one can build a rational model in which a sample at least this apparently overconfident, has a greater than 50% chance (so one can't reject rationality); and one can reject the hypothesis that more than 62% of subjects believe that there is a greater than 80% chance that they are in the top half with confidence levels greater than 99% (the t statistic, for the test that 62% is significantly different from 35%, has 73 degrees of freedom and is -4.9)). At the same time 58% of these subjects are actually in the top half. From

Theorem 2, at least 80% should be. The data passes the first test and fails the second test. (Specifically, fix a rational model in which 35% of a sample of 74 individuals claim to be in the top half with probability at least 80%; the probability that at most 58% of them or less actually score in the top half is at most 0.8%.)

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 fails eight tests at the 10% confidence level. This is consistent with the results of Experiment I, where tests based on Theorem 1 were not stringent enough to detect overconfidence.

4.1 Unskilled and Unaware

Kruger and Dunning (1999) ask subjects to rank themselves on a variety of skills. They find that subjects in the lower quartiles overplace themselves, while subjects in the highest quartile underplace themselves. From this they conclude that overconfidence is the result of subjects who are, in their words, "unskilled and unaware" of their lack of skill. However, the design of their experiment is subject to the criticism of B&D that the subjects' self-rankings could, in fact, be perfectly rational.[18] Nonetheless, our results provide support for their conclusion. In the table below, we show the betting behaviour and placement of subjects as a function of their score on the sample test.

Score on Sample	0	1	2	3	4	5
# with score	3	1	3	14	22	31
Average bet	55	60	26	68	69	71
% in top half	33	0	0	36	41	71

Notably, those who scored five on average predicted they had a 71% chance of being in the top half and 71% ended up in the top half. The overconfidence stems from those who scored three or four. (Only 10% of the subjects score two or less, and ignoring this data has virtually no effect on our results.) While these subjects had around a 40% chance of ending up in the top half, they behaved as if they were as skilled as those who scored a five, also predicting around a 70% chance of ending up in the top half . These subjects appear to have been relatively unskilled, and unaware of it.

[18] In addition, Ackerman, Beier, and Bowen (2002) argue that Kruger and Dunning's finding is actually an example of regression to the mean, though Kruger and Dunning dispute this. See also Krueger and Mueller (2002).

5 Discussion

5.1 Critical Assessment of Data

All our subjects were incentivized to accurately report their beliefs, either implicitly of explicitly. Despite this, there are at least two reasons why the data could be questioned.

1. Subjects may have had goals beyond the maximization of utility derived from their monetary payments. There is evidence that people like to exert control over their situations, and so subjects may have preferred to bet on themselves even in 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). Subjects may also have liked to bet on themselves to present themselves in a positive light. Such motivations would compete with losses in payment. Subjects stood to gain \$10 from winning a bet. While this amount of money is a decent amount for the subject population, it (inevitably) overstates the incentives. For example, in Experiment I, a subject who bets on herself in the strength treatment even though she believes she has only a 30% chance of finishing in the top half, thus implicitly overstating her probability of success by 30%, makes an expected loss of \$3 from this sub-optimal choice.[19] In Experiment II, a subject who overstates her probability of finishing in the top half by 30% makes an expected loss of less than 54 cents (see the appendix). On this accounting, the overconfidence from both experiments may be overstated, and the data from Experiment II may be less reliable than the data from Experiment I. Similar caveats apply to the data from Burks et al. (2010). While they incentivize their subjects to place themselves into their most likely quintile, a subject's loss in payment from stating a higher quintile may be quite small.[20]

2. Although it was 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.[21] For instance, some subjects may have erroneously reasoned that, since stating a higher value ensures that when the randomizing device is used, on average it has a higher probability of succeeding, it is desirable to overstate.[22]

[19] In fact, subjects placed six bets and were rewarded based on test chosen at random. This calculation assumes that subjects consistently overstate across the bets. Overstating only on some bets reduces the expected loss.

[20] Merkle and Weber (2011) do not report the exact formula they use to reward their subjects, so we cannot estimate subjects' losses from overplacing.

[21] 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

[22] See also Plott and Zeiler (2005) whose results show that some of the findings confirming the endowment

Similarly, Merkle and Weber (2011) use the Quadratic Scoring Rule to elicit beliefs – a rule that is also not very intuitive. (Hollard et al. (2010) test the Quadratic Scoring Rule against the Probability Matching Rule for the elicitation of subjective probabilities and find that the Probability Matching Rule provides more accurate beliefs.)

The above two points suggest that our findings of overconfidence may be overstated, more so for Experiment II than Experiment I. Since our two experiments were very similar in terms of the subject population and the quizzes they took, 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 ignore the above two caveats and assume the subjects were expected utility maximizers, 74% of the population believed they had at least a 50% chance of placing in the top half. In Experiment II, 90.5% of subjects reported at least a 50% chance placing in the top half, which is significantly greater than 74%. However, thirteen of these subjects report a probability of exactly 50. If we make a genericity assumption and assume that all the 50's are the result of rounding up (to the nearest even number) in a rational manner, then they should be excluded. Alternatively, if we make a genericity assumption in Experiment I and assume that none of the subjects were indifferent when they made their choices, then we have that all those who bet on themselves strictly preferred this to a 50/50 bet, so that again we should exclude all those who state 50 in Experiment II when making our comparison. If we exclude these thirteen subjects, then we have that 73% of subjects place themselves in the top half, a figure almost identical to the 74% in Experiment I. If we take the middle ground and exclude half the subjects who said 50% in Experiment I we have that 82% place themselves in the top half in Experiment II, and we cannot reject the hypothesis that the two samples 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%.

Similarly, 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%. If we exclude those who say say exactly 60, then the relevant figure for comparison is 61%, which is almost identical to the 64% from the first experiment. If we exclude half of those who say 60%, we have that 66% of those in Experiment II believe there is at least a 60% chance that they are in the top half, and we can't 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 limited evidence that the mechanism used in Experiment II did not, in fact, cause participants to overstate their placement *relative* to the mechanism used in

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.

Experiment I.

5.2 A Reassesment of the Theory

Administering a quiz allows us to incentivize subjects in a way that is difficult to do when asking them about, say, their driving or managerial skills. However, this type of experiment suffers from the fact that the subjects must reflect not only upon their skills but also upon the nature of the quiz they are taking. Moore and Healy (2008) show that when subjects face a quiz that is easier than they expected it to be, even Bayesian reasoning 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 overconfident (see B&D (2011) for a discussion). In order to mitigate this problem, subjects in both our experiments were told how well populations had performed on these quizzes in the past. (As far as we can tell, subjects in Merkle and Weber (2011) or Burks et al. (2011), were not given much information on the difficulty of the quizzes they were taking so that results in these papers may be vulnerable to this critique.)

There is still another issue, which manifests itself when applying Theorem 2. Subjects must consider not only the ease of the test, but its diagnostic value as well. To understand this issue, suppose that a large group of subjects is to take a quiz billed as an "examination of logical reasoning". Suppose that, based on their life experiences to date, 40% of the subjects *rationally* hold the belief that they have at least a 50% chance of ranking in the top 30% on logical ability. Moreover, subjects (have been led to) believe that the examination they are to take is a perfect discriminator of logical ability. Hence, 40% of subjects believe that they have at least a 50% chance of placing in the top 30% on the quiz. This data passes Theorem 1, as it should. Suppose, however, that, contrary to the subjects' belief, the quiz is, in fact, poorly designed and graded so that scores on it are completely arbitrary – all subjects are equally like to score in the top 30%. Then, with a large population, only 30% of the subjects who believe that they have at least a 50% chance of placing in the top 30%, actually place there. The data fails a test based on Theorem 2, even though the subjects are not overconfident. Similarly, if subjects thought the test was one of inductive reasoning, but it was actually one of deductive reasoning, and these skills were imperfectly matched, data might misleadingly fail a test based on Theorem 2.[23] A similar caveat applies to the test used in Burks et al. Such a test is correctly picking up on the fact that subjects have

[23] A similar caveat applies to the test of Burks et al. (2010).

made an error, but the error is one of misunderstanding the nature of the test, not one of overconfidence.

Thus far, we have described an "error" on the part of the subjects who do not properly understand the quizzes they are facing. 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 – unavoidably – not averaging across all experiments. The data may fail the test, but now it is the analyst who is making an error, not the subjects.[24]

6 Conclusion

There is a large body of experiments establishing the better-than-average effect on easy tasks. However, 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. 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 with a good understanding of the nature of the quizzes they took. We have discussed some of the limitations of the approach we have taken, both for our experiments and similar experiments. However, this is not to deny the virtues of this approach. 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. Moreover, questions remain regarding the motives underlying overplacement beliefs, including motives to appear confident, smart, capable, and humble, both to others and to the self.

7 Appendix A: Test items from the two tests

1S) 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

[24] Similarly, with respect to our earlier point on the ease of tests, subjects may correctly understand that some tests are easier than others, while the analyst fails to average across all tests. See B&D for more on this issue.

1M) The Maroons are first in the league and the Browns are fifth while the Blues are between them. If the Grays have more points than the Violets and the Violets are exactly below the Blues then who is second? The Grays

2S) A bridge consists of 10 sections; each section is 2.5 meters long. How far is it from the edge of the bridge to the center? 12.5 m

2M) Five friends share three oranges equally. Each orange contains ten wedges. How many wedges does each friend receive? 6

3S) There are four equally spaced beads on a circle. How many straight lines are needed to connect each bead with every other bead? 6

3M) Fall is to Summer as Monday is to _____? Sunday

4S) HAND is to Glove as HEAD is to _____? Hat

4M) What is the minimum number of toothpicks necessary to spell the word "HAT". (You are not allowed to break or bend any toothpicks, or use one toothpick as a part of more than one letter.) 8

5S) John needs 13 bottles of water from the store. John can only carry 3 at a time. What's the minimum number of trips John needs to make to the store? 5

5M) Milk is to glass as soup is to _____? bowl

6S) LIVED is to DEVIL as 6323 is to _____? 3236

6M) Which number should be next in the sequence: 2, 4, 8, 16, 32, ? 64

7S) If the day before yesterday is two days after Monday then what day is it today? Friday

7M) A rancher is building an open-ended (straight) fence by stringing wire between posts 25 meters apart. If the fence is 100 meters long how many posts should the rancher use? 5

8S) Which number should come next in the series: 3, 9, 6, 12, 9, 15, 12, 18, ? 15

8M) "Meow" is to a cat as "Moo" is to _____? Cow

9S) Which letter logically follows in this sequence: T, Q, N, K, H, ? E

9M) Which word does not belong in the group with the other words? Brown, Black, Broom, Orange, Bread Orange

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

10M) If a woman is 21 and is half the age of her mom, how old will the mom be when the woman is 42? 63

11S) Tiger is to stripes as leopard is to _____? Spots

11M) Which number should come next: 514, 64, 8, 1, 1/8, ? 1/64

12S) Brother is to sister as nephew is to _____? Niece

12M) Which number should come next in this series: 1 - 1 - 2 - 3 - 5 - 8 - 13 - ? 21

13S) Desert is to oasis as ocean is to _____? Island

13M) If 10 missionaries have 3 children each, but only two thirds of the children survive, how many children survive? 20

14S) Kara has $100. She decides to put 20% in savings, donate 20% to a charity, spend 40% on bills, and use 20% for a shopping spree. How much money does she have left over afterwards? $0

14M) Kimberly makes $20 per hour and works for 20 hours each week. How much does she make in a week? 400

15S) How many straight lines are needed to divide a regular hexagon into 6 identical triangles? 3

15M) Which number should come next in this series: 1,4,9,16,25,? 36

16S) What is the average of 12, 6 and 9? 9

16M) DIDIIDID is to 49499494 as DIIDIIDD is to _____? 49949944

17S) There are three 600 ml water bottles. Two are full, the third is 2/3rds full. How much water is there total? 1600ml

17M) If a wood pile contains 30 kilos of wood and 15.5 kilos are burned, how many kilos are left? 14.5

18S) Which letter does not belong in the following series: D - F - H - J - K - N - P - R K

18M) Joe was both 5th highest and 5th lowest in a race. How many people participated? 9

19S) If a certain type of bug lives for only 20 days, how old is the bug when it has lived half of its lifespan? 10 days

19M) PEACH is to HCAEP as 46251 is to _____? 15264

20S) Begin is to began as fight is to _____? Fought

20M) Nurse is to hospital as teacher is to _____? school

8 Appendix B: Proofs

Proof of Theorem 1. Sufficiency. Set $\Theta = \{\theta_l, \theta_h\}$ and $S = \{s_h, s_l\}$, let the joint probability distribution of types and signals be

$$\begin{array}{ccc} & \theta_l & \theta_h \\ s_l & 1-y-(1-q)\,x & y-qx \\ s_h & (1-q)\,x & qx \end{array} \tag{1}$$

Since $0 \leq 1-\frac{y}{q} = 1-y-(1-q)\,\frac{y}{q} \leq 1-y-(1-q)\,x \leq 1$, all the numbers in the matrix are in $[0,1]$. Also, a signal s_h has probability x, and makes the individual believe that there

is a chance of at least q that he is in the top y of the population. If $x = 1$, the proof of sufficiency is finished; if $x < 1$, a signal of s_l has probability $1 - x$ and after it, the belief that the type is θ_h is $\frac{y-qx}{1-x} < q \Leftrightarrow y < q$, so exactly x people believe their type is in the top y with probability at least q.

Necessity. Let $\left(\Theta, p, S, \{f_\theta\}_{\theta\in\Theta}\right)$ be a rationalizing model which rationalizes the data. Since a fraction x believe there is a positive probability that their types are in the top y, there is a $\hat{\theta}$ such that $P\left(\theta \geq \hat{\theta}\right) \leq y$. Let $\overline{\theta} = \min\left\{\hat{\theta} : P\left(\theta \geq \hat{\theta}\right) \leq y\right\}$. Let S_h denote the set of signals such that $P\left(\theta \geq \overline{\theta} \mid s \in S_h\right) \geq q$. We have

$$y \geq P\left(\theta \geq \overline{\theta}\right) = P\left(\theta \geq \overline{\theta} \mid S_h\right) P\left(S_h\right) + P\left(\theta \geq \overline{\theta} \mid S \setminus S_h\right)\left(1 - P\left(S_h\right)\right) \geq P\left(\theta \geq \overline{\theta} \mid S_h\right) P\left(S_h\right) \geq qx.$$

■

Proof of Theorem 2. Sufficiency. Set $\Theta = \{\theta_l, \theta_h\}$ and $S = \{s_h, s_l\}$, and let the joint probability of types and signals be

$$\begin{array}{ccc} & \theta_l & \theta_h \\ s_l & 1 - y + \tilde{x} - x & y - \tilde{x} \\ s_h & x - \tilde{x} & \tilde{x} \end{array} \tag{2}$$

Since $q > y \geq \tilde{x}$, we obtain $1 - y > 1 - q \geq (1 - q)\frac{\tilde{x}}{q} = \frac{\tilde{x}}{q} - \tilde{x} \geq x - \tilde{x}$ so that all numbers in the matrix are in $[0, 1]$. Moreover, a signal of s_h has probability x, and the posterior of θ_h is $\frac{\tilde{x}}{x} \geq q$, so that people observing s_h have a belief of at least q that their type is in the top y of the population. If $x = 1$, the proof of sufficiency is done. If $x < 1$, since $y < q$ and $\tilde{x} \geq qx$, a signal of s_l assigns a probability $\frac{y-\tilde{x}}{1-x} < q$ to θ_h and therefore those who observe signal s_l do not declare their type in the top y with probability at least q.

Necessity. Let $\left(\Theta, p, S, \{f_\theta\}_{\theta\in\Theta}\right)$ be any rationalizing model which rationalizes the data. Since a fraction x believe there is a positive probability that their types are in the top y, there is a $\hat{\theta}$ such that $P\left(\theta \geq \hat{\theta}\right) \leq y$. Let $\overline{\theta} = \min\left\{\hat{\theta} : P\left(\theta \geq \hat{\theta}\right) \leq y\right\}$. Let S_h denote the set of signals such that $P\left(\theta \geq \overline{\theta} \mid s \in S_h\right) \geq q$. We have

$$\tilde{x} = P\left(\theta \geq \overline{\theta} \& S_h\right) = P\left(\theta \geq \overline{\theta} \mid S_h\right) P\left(S_h\right) \geq qx.$$

■

The next 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 suppose 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. ■

Our mechanism in Experiment II can be summarized as follows: say a number n between 1 and 50; the computer selects a number $x \in [1, 50] \cap \mathbf{N}$. If $x \geq n$, you win \$10 with probability $2x$ (we draw a bingo ball, and if it is lower than $2x$ you win \$10); if $x < n$, you win \$10 if your score is in the top half.

What is the value of reporting n when the belief is b?

$$v(b, n) = \sum_{x=1}^{x=n-1} \frac{1}{50}\frac{b}{50}10 + \sum_{x=n}^{x=50} \frac{1}{50}\frac{x}{50}10 = \frac{1}{500}n - \frac{1}{250}b + \frac{1}{250}bn - \frac{1}{500}n^2 + \frac{51}{10}$$

Since the mechanism elicits the smallest even number larger than the individual's belief, suppose the individual must round up his belief b to $b + r$, for $r < 2$. Then, if the individual overstates changes his optimal bet by 30%, it means he is declaring an $n = b + 15 + r$, so

$$v(b, b + r) - v(b, b + 15 + r) = \frac{21 + 3r}{50} < 54 \text{ cents.}$$

9 Appendix C: Tests in Experiment II

Table 1 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 71.6% 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 a $t-$statistic and a p-value for a test that the data comes from a population in which $\frac{y}{q}$ or more of the population think they are in the top y with probability at least q, the sixth entry indicates that 43.2% 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 seventh entry indicates the probability that in a particular rational model (the one that maximizes the chance that, in a sample in which x claim to be in the top y, a fraction $\tilde{x}$ will claim to be in the top y and score there), in a sample in which x claim to be in the top y, a fraction $\tilde{x}$ or less claim to be in the top y and score there.

For each row in the table, three tests can be conducted, in principle. Based on Theorem 1, if it were the case that $qx > y$, one could ask whether this difference is statistically significant using as a null rationality (first test), or using as a null a particular form of irrationality

(second test). It is easy to check that if $qx \leq y$, then with $\Theta = \{\theta_l, \theta_h\}$, $S = \{s_l, s_h\}$, $p(\theta_h) = \frac{1}{2}$ and likelihood functions given by

$$\begin{array}{ccc} & \theta_l & \theta_h \\ s_l & 1-(1-q)\,x & 1-qx \\ s_h & (1-q)\,x & qx \end{array} \tag{3}$$

the likelihood of a sample in which at least $x\%$ of the population declares to be in the top $y = \frac{1}{2}$ of the population is greater than 50%. Since in every case $y > qx$, rationality can never be rejected. The table then presents, for each line, a test of the hypothesis that the sample comes from a population in which a proportion greater than $\frac{y}{q}$ has a belief that they are in the top y with probability at least q (i.e. the null is a particular form of irrationality). The fifth column presents the $t-$statistic for that test, and the p value (for a distribution with 73 degrees of freedom). This form of irrationality is rejected in every case: in two cases the p value is 1.5%, in the rest it is less than 1%.

The third set of tests is as follows. Given $(y, q, x, \tilde{x})$, set $\Theta = \{\theta_l, \theta_h\}$, $S = \{s_l, s_h\}$, $p(\theta_h) = \frac{1}{2}$. We then choose the likelihood functions to maximize the probability that a proportion x will observe signal s_h; this yields

$$\frac{f_{\theta_l}(s_h) + f_{\theta_h}(s_h)}{2} = x.$$

In order for the conditional of θ_h given s_h to be at least q we need

$$p(h \mid s_h) = \frac{f_{\theta_h}(s_h)\frac{1}{2}}{f_{\theta_l}(s_h)\frac{1}{2} + f_{\theta_h}(s_h)\frac{1}{2}} \geq q \Leftrightarrow f_{\theta_h}(s_h) \geq \frac{q}{1-q} f_{\theta_l}(s_h).$$

- It is easy to check that if $\tilde{x} \geq qx$ then the following likelihood functions

$$\begin{array}{ccc} & \theta_l & \theta_h \\ s_l & 1-2(x-\tilde{x}) & 1-2\tilde{x} \\ s_h & 2(x-\tilde{x}) & 2\tilde{x} \end{array} \tag{4}$$

  maximize the likelihood that in a sample in which x claim to be in the top y, a fraction $\tilde{x}$ will claim to be in the top y and score there. Recall that because $\tilde{x}$ is the fraction of people who claim to be in the top half, and are actually in the top half, we obtain $\frac{1}{2} \geq \tilde{x}$; this implies that f_{θ_h} is indeed a probability distribution. Also, since $\frac{1}{2} = y \leq q$ in the table below, we have $(1-q)\,x \leq \frac{x}{2} \leq \frac{1}{2} \Rightarrow x \leq qx + \frac{1}{2} \leq \tilde{x} + \frac{1}{2}$, which implies that f_{θ_l} is also a probability distribution. With the model in (4) the likelihood that in a sample in which x claim to be in the top y, a fraction $\tilde{x}$ or less claim to be in the top y and score there is always greater than 50%, as reported in the last column of Table 1 (see for example the first three rows of the table).

- Suppose instead that $\tilde{x} < qx$. The likelihood functions that maximize the probability that, conditional on x claiming to be in the top y, a fraction $\tilde{x}$ will claim to be in the top y and score there are given by (3). The likelihood that, in a sample in which x claim to be in the top y, a fraction $\tilde{x}$ or less claim to be in the top y and score there are also reported in the last column of Table 1.

Table 1

		Test based on Theorem 1			Test based on Theorem 2	
y	q	x	qx	$t-$ stat; p value	$\tilde{x}$	Likelihood
50%	50%	$\frac{67}{74} = 90.5\%$	45.3%	$-\infty; 0$	$\frac{35}{74} = 47.3\%$	$> 50\%$
50%	58%	$\frac{54}{74} = 73.0\%$	42.3%	$-2.55; 0.6\%$	$\frac{33}{74} = 44.6\%$	$> 50\%$
50%	60%	$\frac{53}{74} = 71.6\%$	43.0%	$-2.22; 1.5\%$	$\frac{32}{74} = 43.2\%$	$> 50\%$
50%	66%	$\frac{45}{74} = 60.8\%$	40.1%	$-2.62; 0.5\%$	$\frac{27}{74} = 36.5\%$	24.2%
50%	68%	$\frac{44}{74} = 59.4\%$	40.4%	$-2.45; 1.5\%$	$\frac{26}{74} = 35.1\%$	13.5%
50%	70%	$\frac{41}{74} = 55.4\%$	38.8%	$-2.75; 0.4\%$	$\frac{24}{74} = 32.4\%$	7.9%
50%	72%	$\frac{32}{74} = 43.2\%$	30.4%	$-4.5; < 0.1\%$		
50%	74%	$\frac{31}{74} = 41.9\%$	31.0%	$-4.5; < 0.1\%$	$\frac{18}{74} = 24.3\%$	3.9%
50%	76%	$\frac{30}{74} = 40.5\%$	30.8%	$-4.4; < 0.1\%$	$\frac{17}{74} = 23.0\%$	1.5%
50%	78%	$\frac{27}{74} = 36.5\%$	28.5%	$-4.9; < 0.1\%$	$\frac{16}{74} = 21.6\%$	2.2%
50%	80%	$\frac{26}{74} = 35.1\%$	28.1%	$-4.9; < 0.1\%$	$\frac{15}{74} = 20.3\%$	0.8%
50%	84%	$\frac{17}{74} = 23.0\%$	19.3%	$-7.4; < 0.1\%$		
50%	86%	$\frac{15}{74} = 20.3\%$	17.4%	$-8.0; < 0.1\%$	$\frac{11}{74} = 14.9\%$	14.8%
50%	88%	$\frac{14}{74} = 18.9\%$	16.7%	$-8.3; < 0.1\%$	$\frac{10}{74} = 13.5\%$	7.7%
50%	90%	$\frac{13}{74} = 17.6\%$	15.8%	$-8.5; < 0.1\%$	$\frac{9}{74} = 12.2\%$	3.4%
50%	92%	$\frac{6}{74} = 8.1\%$	7.5%	$-14; < 0.1\%$		
50%	94%	$\frac{5}{74} = 6.8\%$	6.4%	$-15; < 0.1\%$	$\frac{4}{74} = 5.4\%$	3.2%
50%	96%	$\frac{4}{74} = 5.4\%$	5.2%	$-17; < 0.1\%$		
50%	98%	$\frac{3}{74} = 4.1\%$	3.9%	$-20; < 0.1\%$	$\frac{3}{74} = 4.1\%$	15.1%
50%	100%	$\frac{2}{74} = 2.7\%$	2.7%	$-25; < 0.1\%$	$\frac{2}{74} = 2.7\%$	$> 50\%$

The value $\tilde{x}$ is calculated as follows (a similar calculation was made for Experiment I). A total of 27 people scored 19 or 20, and 13 score 18 (the median score). Hence, if a person scores 18, his chance of being in the top half of test takers is $\frac{74/2-27}{13} = \frac{10}{13}$. For each line in the table above we determine how many of the individuals claimed to be in the top 50% with probability greater than q, score 19 or 20; to those we add a proportion $\frac{10}{13}$ of those who claimed to be in the top 50% with probability greater than q and scored 18.

References

Ackerman, P. L., M. E. Beier and K.R. Bowen (2002), "What we really know about our abilities and our knowledge," *Personality and Individual Differences*, **33(4)**, 587-605.
Amabile, T. (1983) "The social psychology of creativity: A componential conceptualization," *Journal of Personality and Social Psychology*, **45(2)**, 357-76.
Anderson, C., S. Srivastava, J.S. Beer, S.E. Spataro and J.A. Chatman (2006), "Knowing your place: Self-perceptions of status in face-to-face groups," *Journal of Personality and Social Psychology*, **91(6)**, 1094-1110.
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
Armor, D. A., C. Massey, and A.M. Sackett (in press), "Prescribed optimism: Is it right to be wrong about the future?" *Psychological Science*.
Barber, B. and T. Odean (2001), "Boys Will Be Boys: Gender, Overconfidence, And Common Stock Investment," *Quarterly Journal of Economics*, 116(1), 261-92.
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, J-P. and J. Dubra (2008), "Overconfidence?" at SSRN, http://ssrn.com/abstract=1088746
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 (2010), "Overconfidence is a Social Signalling Bias," IZA Discussion Paper 4840.
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 (2008), "Rational Expectations of Own Performance: An Experimental Study," forthcoming *Economic Journal*.
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., J.A. Meyerowitz and A.D. Holzberg (1989), "Ambiguity and self-evaluation:

The role of idiosyncratic trait definitions in self-serving assessments of ability," *Journal of Personality and Social Psychology*, **57(6)**, 1082-90.
Fang, H. and G. Moscarini, (2005) "Morale Hazard," *Journal of Monetary Economics*, **52(4)**, 749-777.
Festinger, L. (1954) "A Theory of Social Comparison Processes," *Human Relations*, **7(2)**, 117-140.
Garcia, D., F. Sangiorgi and B. Urosevic, (2007), "Overconfidence and Market Efficiency with Heterogeneous Agents," *Journal Economic Theory*, **30(2)**, 313-36.
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. (1980), "Bayes' rule as a descriptive model: The representative heuristic," *Quarterly Journal of Economics*, **95**, 537-557.
Grether, D. M. (1981) "Financial Incentive Effects and Individual Decision Making," Social Sciences working paper 401, Cal. Tech.
Grether, D. M. (1990), "Testing Bayes rule and the representativeness heuristic: Some experimental evidence," *Journal of Economic Behavior and Organization*, **17**, 31-57.
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.
Hollard, G., S. Massoni and J-C Vergnaud, (2010), "Subjective beliefs formation and elicitation rules: experimental evidence," CES Sorbonne WP 2010.88.
Kahneman, D. and A. Tversky (1972), "Subjective probability: A judgment of representativeness," *Cognitive Psychology*, **3(3)**, 430-454.
Karni, E. (2009), "A Mechanism for Eliciting Probabilities," *Econometrica*, **77(2)**, 603–6.
Kőszegi, B., (2006), "Ego Utility, Overconfidence, and Task Choice," *Journal of the European Economic Association*, **4(4)**, 673-707.
Krizan, Z., & Windschitl, P. D. (2007), "The influence of outcome desirability on optimism," *Psychological Bulletin*, 133(1), 95-121.
Krueger J. and RA Mueller (2002), "Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance," *Journal of Personality and Social Psychology*, **82(2)**, 180-8.
Kruger, J. (1999), "Lake Wobegon Be Gone! The "Below-Average Effect" and the Egocentric Nature of Comparative Ability Judgements", *Journal of Personality and Social Psychology*,

77(2), 221-232.

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.

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.

McGraw, A. P., B.A. Mellers and I. Ritov (2004), "The affective costs of overconfidence," *Journal of Behavioral Decision Making*, **17(4)**, 281-295.

McKelvey, R. D. and T. Page (1990), "Public and private information: An experimental study of information pooling," *Econometrica*, **58**, 1321-39.

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

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.

Sonnemans, J. and T. Offerman (2001), "Is the Quadratic Scoring Rule really incentive compatible?"

Stone, D. N. (1994),"Overconfidence in initial self-efficacy judgments: Effects on decision processes and performance," *Organizational Behavior and Human Decision Processes*, **59(3)**, 452-74.

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

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.

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