A proper test of overconfidence

Benoît, Jean-Pierre; Dubra, Juan and Moore, Don
London Business School, Universidad de Montevideo, Tepper School of Business, Carnegie Mellon University

05. December 2008

A Proper Test of Overconfidence*

Jean-Pierre Benoît
London Business School

Juan Dubra†
Universidad de Montevideo

Don Moore
Tepper School of Business, Carnegie Mellon University.

Abstract

In this paper we conduct two proper tests of overconfidence. We reject the hypothesis "the data cannot be generated by a rational model" in both experiments.

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 purports to find that people are generally overconfident. In particular, a better-than-average effect in which a majority of people claim to be superior to the average person has been noted for a wide range of skills, from driving, to spoken expression, to the ability to get along with others, to test taking on simple tests.[1] The literature generally accepts that this better-than-average effect is indicative of inflated self-assessments. However, Benoît and Dubra (2008) (henceforth B&D) have recently questioned this stance. They show that the better-than-average data merely has the *appearance* of overconfidence, but does not indicate *true* overconfidence, which carries with it an implication that people have made some kind of error in their self-evaluations. Because of this reason, almost none of the existing experimental literature on relative overconfidence can actually claim to have found overconfidence. In fact, most of the experiments by their very design do not even have the potential of showing overconfidence. In this paper, we report on an experiment designed to provide a proper test of overconfidence. The issue of whether people are in fact overconfident is of paramount importance for economics as it determines the

*We thank Uriel Haran for his help with the experiment.

†email: dubraj@um.edu.uy

[1] See Benoît and Dubra (2008) for a review. 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.

equilibrium outcomes in almost every market.[2] Having a correct test of overconfidence is then relevant. The following example taken directly from B&D illustrates the basic idea why previous tests have been misspecified.

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?

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

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

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 of people could rank themselves as being in the top half of the population without any overconfidence being implied. Therefore, any experiment designed just to test for a better-than-average-effect cannot possibly show overconfidence.

We conduct a test that has the potential to reveal that people are not making rational assessments of their abilities. The experiment is based upon the theory developed in B&D, which we briefly review in Section 2. Although the subjects in our experiment also give

[2] Papers on overconfidence in economics include Camerer and Lovallo (1999) who study experimentally entry in an industry, Fang and Moscarini (2005) analyzing the effect of overconfidence on optimal wage setting, Garcia, Sangiorgi and Urosevic (2007) who characterize the efficiency consequences of overconfidence in information acquisition in financial markets, Hoelzl and Rustichini (2005) who test for overconfidence, Kőszegi (2006) who studies how overconfidence affects how people choose tasks or careers, Menkhoff et al. (2006) who analyze the effect of overconfidence on herding by fund managers, Noth and Weber (2003), Sandroni and Squintani (2008), Van den Steen (2004) and Zábojník (2004). In finance, recent papers include Barber and Odean (2001), Biais et al. (2005), 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), Wang (2001).

the superficial appearance of being overconfident, we do not find any evidence that they are in fact overconfident. To our knowledge, the only other experiment that conducts a proper test of a better-than-average bias, is Clark and Friesen (2008), and they also do not find such a bias. In fact, the experiment of Camerer and Lovallo (1999), when correctly interpreted, also provides such a test. More concretely, Camerer and Lovallo (1999) claim to find overconfidence because their subjects enter an industry more when their payoff depends on their score on a trivia quiz than when it depends on a random draw. B&D argue that that test is misspecified because that is not the right statistic to look at; rather, one should look at profits made by entrants, and profits are positive in the treatment without self selection, indicating that people are in fact *not* overconfident (see B&D for more on this issue).

The most common type of experiment in this field asks subjects how they rank compared to others. For instance, Weinstein (1980) asks students to compare themselves to the average student on a variety of attributes including their chances getting a good job offer before graduation and their chances of developing a drinking problem. Similarly, Svenson (1981) asks subjects in a room to estimate how their driving compares to the other subjects in the room, and to make an estimate of the form "I drive better than x% of the people in this room".

There are at least four criticisms that can be made of this type of experiment:

1. Participants have no material incentive to answer the question accurately.

2. It may be unclear to the subjects what is meant by an "average" student. In particular, should the average be interpreted as the mean or median (or something else still)?

3. Subjects may be uncertain of their *own* skill levels, making the meaning of their answers unclear.

4. No attention is paid to the degree of confidence that the subjects have in their self-placements.

The first two criticisms are quite familiar, so let us turn to the last 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, for the sake of argument, 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, we have no way of knowing his degree of confidence when he utters a statement like "I believe I have a higher IQ than the average person". As we will discuss, both these ambiguities have important implications. Note, however, that if, as a matter of fact, subjects have very tight estimates of their types then both these issues become moot – the means and medians will be about the same and subjects will be almost 100% confident in their self-placements. In addition to testing for overconfidence we test the hypothesis that subjects do not have very tight estimates of their types. Note that in the previous driving example, $\frac{3}{5}$ of the drivers

believe that their mean abilities and median abilities are better than average, justifying their overconfident seeming answers. At the same time, each of these drivers thinks there is a $\frac{4}{9}$ chance that he is not above average, and even a $\frac{1}{9}$ chance that he is below average.

Even if we grant that subjects with no material incentive respond to questionnaires as accurately as possible, so that point 1) above is not an issue, an experiment that fails to pay attention to *any* one of the remaining points will fail as a test of overconfidence, as we show in the following section.

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 relative self-assessments. 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 not more than 50% 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. Obviously, it is important to have a proper theoretical framework for discussing overconfidence.

Clearly, 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.[3] Thus, B&D proposes that data be called *overconfident* only if it cannot be obtained from a population which derives its beliefs in a fully rational and consistent manner. A fairly standard model for a population deriving its beliefs in such a manner is as follows:

Definition 1 *A **signalling structure** is a triplet $\sigma = (S, \Theta, f)$, where S is a set of signals, $\Theta \subset \mathbf{R}$ is a type space, and $f = \{f_\theta\}_{\theta \in \Theta}$ is a collection of probability distributions over S.*

Definition 2 *A **signalling model** consists of a population of individuals and a signalling structure $\sigma = (S, \Theta, f)$ such that:*

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

i) In period 0, nature picks a type $\theta \in \Theta$ for each individual, resulting in some distribution p; initially, each person's belief about her own type is given by this distribution.
ii) In period 1, an individual of type θ receives a signal $s \in S$ according to the probability distribution f_θ; each person updates her initial belief using Bayes' rule.

- Throughout this paper we assume that higher types are more skillful.

A person of type t is said to be in the top x of a population if the fraction of people whose type is greater than or equal to t is at most x. Thus, in a population of 100 people *at most* 30 can be in the top 30%.

Definition 3 *Suppose that a fraction y of a population of N people believe that there is a probability q that their type is in the top x of the population. These beliefs can be* ***rationalized*** *if there is a signalling model with N individuals in which the expected fraction of people who will have these beliefs after updating is y.*

Notice that by asking that y be the *expected* fraction of people who will hold the particular beliefs, the definition is demanding: Data cannot be rationalized simply because it is possible that it could arise in a stochastic environment. If the data from an experiment can be rationalized, there is no reason to call it overconfident.

The following Theorem, taken from B&D, provides the basis for our tests of overconfidence.

Theorem 1 *Consider a population of N people and two integers $0 \leq m \leq N$ and $1 \leq r \leq N$. Suppose a fraction $y = \frac{m}{N}$ of the population believe that there is a probability at least q that their types are in the top $x = \frac{r}{N}$ of the population. These beliefs can be rationalized if and only if $qy \leq x$.*

The following example illustrates the Theorem. Consider ten people who are to take a math test. First suppose that seven of them believe that there is at least a $\frac{1}{2}$ probability that their type is in the top 3 (so that $qy > x$). If this belief were rational, then on average at least $\frac{1}{2} \times \frac{7}{10} \times 100\% = 35\%$ of the population would be in the top 30%, a clear absurdity. On the other hand, suppose instead that $\frac{6}{10}$ of the people believe that there is at least a $\frac{1}{2}$ probability that their type is in the top 30%. How could these beliefs rationally arise? One simple way is as follows. Before the test, a brief conversation reveals that six of them have an advanced degree in mathematics, whereas the remaining four have only high school mathematics. With no further information, the six can rationally believe they will place in the top six, with the precise order being uniformly random. Hence each of the six believes there is a $\frac{1}{2}$ chance he or she will place in the top 30%.

Armed with Theorem 1, we are in a position to better appreciate the four criticisms of prior experiments made in the introduction.

Consider a person who is given the choice between a 50% chance at a prize, and the prize if she places in the top half of a subject pool on a test (as in Hoelzl and Rustichini, 2005). The person has been incentivized and the meaning of "average" is irrelevant , so Criticism 1 and 2 are irrelevant. Suppose the person strictly prefers the prize based on her test placement. The meaning of her preference is clear –she believes that there is more than

a 50% chance that she places in the top half – so that Criticism 3 does not apply. However, the strength of this belief – exactly how much more than 50% – is unclear, so that Criticism 4 applies. Theorem 1 tells us that almost everybody could rationally prefer the placement alternative, rendering the experiment useless as a test for overconfidence.[4]

Svenson (1981) finds that 46% of (American) subjects in his experiment claim to be in the top 20% of subjects in their driving skill level. His subjects are not incentivized. More importantly, even granting the veracity of their answers, the meaning of these claims is unclear (Criticism 3). If the subjects, who presumably are uncertain of exactly how skillful they are as drivers, are answering based upon their self-beliefs about their median type, then Theorem 1 shows that the subjects are displaying overconfidence. However, if the subjects are answering based upon their self-beliefs about their mean type, then Theorem 4 in B&D shows that their answers are consistent with purely rational self-assessments.[5]

3 The experiment

We are interested in the extent to which previous findings of apparent overconfidence could be shown to be actual overconfidence. Previous experimental work and the theory in B&D show that populations exhibit the better-than-average effect more on easy tasks than difficult ones.[6] Accordingly, we gave our subjects an easy test and motivated them.

Subjects were 134 individuals recruited through the web site of the Center for Behavioral Decision Research at Carnegie Mellon University <http://cbdr.cmu.edu/experiments/>. We will report the data for the 129 subjects who had 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 would depend on skill). So, if anything, our results are biased towards finding overconfidence (but we don't).

Subjects had a mean age of 25 years (SD = 6.4) and 42 percent of them were male. All subjects took a 20-item quiz of math and logic puzzles. 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. The three pairs of bets are listed below.

[4] Even everybody (as opposed to almost everybody) strictly prefering the placement bet is consistent with rationality, given an inevitable "sampling error" due to the finite population.

[5] See B&D for a more detailed discussion of the issue.

[6] The theory in Healy and Moore (2007) also predicts that an easy test should yield more overconfident looking data.

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 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.

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. 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 would 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 the 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.

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.[7]

Then subjects chose between the skill and chance options for each of the three bets again.

Subjects then answered a series of questions (shown in Appendix B) 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.

4 The data

Before taking the test, each subject was presented with the three previously listed groups of choices. The order in which subjects were presented with these choices was randomized among subjects. The choices can be summarized as:

1. **Benchmark Choice**: 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 Choice**: 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 Choice**: 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.

There are 5 variables, none of which had any effect on the choice behavior of subjects (or their scores).

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.

[7] In the design of Hoelzl and Rustichini everybody in a group was paid according to the same criterion (skill or chance) after the individuals had voted for which one it would be. In order to avoid the multiplicity of equilibria in the voting stage, we paid each subject according to his choice. We then introduced the payment per correct answer in order to avoid shirking by those who had chosen the random bets.

- 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.[8] 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 Choice, 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 Choice, 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 Choice, 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 Choice, the population displays apparent overconfidence: 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 (or more than 50% chance if we interpret their preferences as being strict).

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 Theorem 1 shows.

Thus, an immediate question of interest is how strong is a subject's belief that she is in the top half. Imagine that we gave the prior literature on overconfidence the benefit of the doubt and interpreted, to the literature's advantage, that the general acceptance of the better-than-average effect as evidence of overconfidence was the consequence of the shared and unstated belief that people are certain (or almost certain) of their types. This assumption, however, would conflict with much of the psychology and behavioral economics literatures that show (or sometimes assume) that people are often engaged in a continual process of updating.

In the behavioral economics literature some influential papers are based on the presumption that "learning about oneself is an ongoing process" (Benabou and Tirole (2002) ; see also Kőszegi, 2006). There are also a few fields within the psychology literature that stress that subjects are uncertain of their types. As a first example of these strands, Festinger's (1954) social comparison theory, the insights of which later permeated to several other fields, begins with his Hypothesis I that "there exists, in the human organism, a drive to evaluate

[8] Other modelings are possible, however, and the choice of modelings is not without consequence.

his opinions and abilities". As a justification for why people wouldn't know their types, he writes "the unavailability of the opportunity for... clear testing and the vague and multipurpose use of various abilities generally make... a clear objective test not feasible or not useful. For example, how does one decide how intelligent one is?" A consequence of this uncertainty is that people change their estimates of their abilities when they receive feedback. In general, in "the absence of both a physical and a social comparison, subjective evaluations of opinions and abilities are unstable". He then adds, "even after a person has had a good deal of experience at a task, the evaluation of what is good performance continues to fluctuate".

The ideas put forth by Festinger, have led psychologists to study the "drive to learn one's abilities" (see, Trope (1975) and Trope and Brickman (1975) *inter alia*). Another example of an influential theory that postulates that people learn about their abilities and about themselves is Bem's (1967) self-perception theory. His central argument is that individuals acquire self-knowledge by observing their own behaviour just as an observer would, especially when "internal cues are weak, ambiguous or uninterpretable". Finally, Amabile's (1983) hypothesis is that if an individual is told that he is creative, his estimate of his creativity changes.

So how strong is a subject's belief that she is in the top half? 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.[9] 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 (2008), and of Healy and Moore (2008), that people are uncertain of their types. In particular, it shows that the prior work on overconfidence cannot be justified by a presumption that people are certain, or nearly certain, of their types. Presumably, if we had asked people to vote for their placement versus a 70% random draw we would have found even more people defecting from the placement option.

We turn now to the question of overconfidence. As noted, Theorem 1 indicates that the Benchmark Choice cannot show overconfidence since every one could prefer the placement option even in a rational population. However, the Strength Choice and High Placement Choice do have the potential to show overconfidence.

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 83.3% vote for the skill bet in the Strength pair of bets. In fact, only 51.9% and not 60% vote for the skill bet in the High Placement pair of bets; the probability that a sample where 51.9% would vote for skill when in the population the proportion is really 60% is 3%. Put differently, 51.9% is different from 60% at the 3% significance level; the probability that this sample comes from an overconfident population in which 60% would vote for the skill placement is 3%. Also, only 64.3% and not 83.3% choose the skill bet in the Strength pair; the probability that a sample where 64.3% would vote for skill when in the population the proportion is really 83.3% is less than 1%. Put differently, 64.3% is different from 83.3% at significance levels lower than 1%; the probability that this sample comes from an overconfident population in which 83.3% would vote for the skill placement is less than 1%.

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

Thus, we do not find evidence of overconfidence.[10]

4.1 A Single Model

Theorem 1 indicates that the results from our three questions can all be generated in a rational fashion. More precisely, the theorem tells us that the data from these three choices can be rationalized by three different rational models (three populations, three signalling structures, etc). However, our data comes from a single subject pool in a single experiment. We now show by construction that the data can also be generated by a single experiment in which all the participants are fully rational.

There are twenty-one possible scores in our experiment, and so there are twenty-one types. Subjects receive signals of their types. (These signals are their life experiences and the sample test they are shown.) Given the nature of the experiment, the simplest model to generate the data is one in which the population divides into three "equivalence classes". Types in the lowest equivalence class, θ_l, have a score in the bottom 50% of subjects; types in the middle equivalence class, θ_m, have a score in-between the bottom 50% and the top 30% of subjects; types in the highest class have a score in the top 30%.

Suppose that each type in a given equivalence class receives one of fours signals s_1, s_2, s_3, s_4 according to the same probability distribution. The joint probability distribution of types and signals is

	θ_l	θ_m	θ_h	Marginal
s_1	.2599081	.000087	0000049	26%
s_2	.0499	.0393	.0108	10%
s_3	.051987	.043823	.03419	13%
s_4	.1382049	.11679	.2550051	51%
Marginal	$\frac{1}{2}$	$\frac{1}{5}$	$\frac{3}{10}$	

The numbers in the above chart are not particularly "nice" as they must be chosen to fit the data. Importantly, however, the signalling structure itself is nice in that it satisfies the monotone likelihood ratio property (for $\theta' > \theta$, we have that $\Pr_{\theta'}(s_i) / \Pr_{\theta}(s_i)$ is increasing in s_i).

[10] We also conducted a variation of the experiment which was identical in all respects, except that subjects were asked to vote after taking the test. In this treatment players are essentially guessing at how many questions they answered correctly after the fact. Indeed, in the feedback treatment subjects are told the test scores against which they are competing, so the only question is how their own scores compare to these given scores. The subjects did not seem particularly adept at this task and exhibited some degree of overestimation of their scores (this is not about better than average). Specifically, 77.5% vote for skill in the Benchmark Treatment, 65.1% vote for skill in the High Placement treatment, while 73.6% vote for skill in the Strength Treatment. However, this seems to be a different type of enquiry than that which is usually addressed in the better-than-average literature.

The following table shows $\Pr(\theta_i|s_j)$

	θ_l	θ_m	θ_h
s_1	0.99965	.00033462	.000018846
s_2	0.499	0.393	0.108
s_3	0.399 9	0.337 1	0.263
s_4	0.270 99	0.229	0.50001

Thus, a person who sees the signal, say, s_4, believes she has just above a 27% chance of placing in the bottom 50%, just below a 23% chance of placing higher than the bottom 50% but lower than the top 30%, and just above a 50% of placing in the top 50%. Such a person will always vote for the placement option rather one of the random choices, since there is a 73% chance that she places in the top 50% and over a 50% chance that she places in the top 30%. The following table indicates how people who receive the different signals should vote:

	Pr	top half vs 50%	top 30% vs 50%	top half vs 60%
s_1	26%	Random	Random	Random
s_2	10%	Placement	Random	Random
s_3	13%	Placement	Random	Placement
s_4	51%	Placement	Placement	Placement
	Placement Total	74%	51%	64%

As the bottom row of the table shows, this signalling model generates the data found in our experiment.

5 Conclusion

As in much previous experimental work, we find a better-than-average effect among our subjects. Since the task we assigned the subjects was an easy one, the theory in B&D led us to expect this finidng. In contrast to previous work, we push further to see if the subjects exhibit behaviour that is truly indicative of overconfidence. We do not find such evidence. At the same time, we find evidence that subjects are uncertain of their own types.

Our experiment can be viewed as a test of the null hypothesis that people are behaving rationally (and are not overconfident). We cannot reject that hypothesis. Of course, this is not to say that we can rule out the hypothesis that people are overconfident, either. In fact, by their very design these types of experiment are ill-suited to rule out overconfident, or underconfident, behaviour. To understand this claim, suppose that ten subjects are to be given a Japanese vocabulary test and that nine of them have absolutely no knowledge of Japanese, while the tenth is Japanese. The nine subjects, who will answer questions randomly, each have about a $\frac{4}{9}$ chance of finishing in the top half while the Japanese subject will almost certainly finish in the top half. If the subjects are behaving rationally, only 10% of the people should prefer betting that they place in the top half rather than accepting a 50% chance at a prize. Therefore, if 30% vote for the placement option, the subjects, as a whole, are overconfident even though they naively appear to be underconfident.

6 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

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

7 Appendix B: The post-test questionnaire

1. How do you think you did on the test? Please estimate your score:

I estimate that I answered _____ of the 20 questions correctly.

2. You probably aren't completely sure of exactly how you scored. Please write down a range of scores below, such that you are 90% sure that your actual score falls somewhere in the range:

I am 90% sure that I have answered between _____ and _____ questions correctly.

3. Please estimate your percentile ranking. In other words, what percentage of other test takers do you think had scores lower than yours?

I estimate that ____% of all participants had scores lower than mine on the test.

4. You probably aren't completely sure if this percentile ranking either. Please write down a range of percentile rankings, such that you are 90% sure that your actual percentile ranking falls somewhere in the range:

I am 90% sure that between ____% and ____% of other test-takers had scores below mine.

5. How pleasant was the experiment for you?

1—2—3—4—5—6—7—8—9—10

Very unpleasant Very pleasant

6. How important was it for you to answer the questions correctly?

1—2—3—4—5—6—7—8—9—10

Not important Very important

7. How important was it for you to estimate your performance accurately?

1—2—3—4—5—6—7—8—9—10

Not important Very important

8. How important was it for you to answer more questions correctly than other participants?

1—2—3—4—5—6—7—8—9—10

Not important Very important

9. To what extent did you invest effort in solving the questions?

1—2—3—4—5—6—7—8—9—10

No effort Extreme effort

10. How did your performance on this test compare with your general ability on tests of this kind?

(1) I did much worse on this test than I ought to have, given my abilities
(2) I did a little worse on this test than I ought to have, given my abilities
(3) My performance accurately reflects my ability
(4) I did a little better on this test than I ought to have, given my abilities
(5) I did much better on this test than I ought to have, given my abilities

11. Have you ever taken an IQ test before? (please circle one) Yes No

12. If yes, what was your score?

(1) Under 80
(2) 80-90
(3) 90-100
(4) 100-110
(5) 110-120
(6) Over 120
(7) Don't know / don't remember

References

Alicke, M. D., M. L. Klotz, D. L. Breitenbecher, T. J. Yurak, and D.S. Vredenburg, (1995), "Personal contact, individuation, and the better-than-average effect," *Journal of Personality and Social Psychology*, **68(5)**, 804-825.

Amabile, T. "The social psychology of creativity: A componential conceptualization," *Journal of Personality and Social Psychology*, **45(2)**, 357-76.

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.

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*.

Festinger, L. (1954) "A Theory of Social Comparison Processes," *Human Relations*, **7(2)**, 117-140.

Healy, P.J. and D. Moore, (2007), "Bayesian Overconfidence," SSRN working paper: http://ssrn.com/abstra
Hoelzl, E. and A. Rustichini, (2005), "Overconfident: do you put your money on it?" the *Economic Journal*, **115**, pp. 305-18.
Kőszegi, B., (2006), "Ego Utility, Overconfidence, and Task Choice," *Journal of the European Economic Association*, **4(4)**, 673-707.
Svenson, O., (1981), "Are we all less risky and more skillful than our fellow drivers?" *Acta Psychologica*, **94**, pp 143-148.
Trope, Y. (1975), "Seeking information about one's own ability as a determinant of choice among tasks," *Journal of Personality and Social Psychology*, **32(6)**, 1004-13.
Trope, Y. and P. Brickman (1975), "Difficulty and diagnosticity as determinants of choice among tasks," *Journal of Personality and Social Psychology*, **31(5)**, 918-25.
Weinstein, N. (1980), "Unrealistic Optimism about Future Life Events," *Journal of Personality and Social Psychology*, **39(5)**, 806-20.