The Economic Journal, **115** (*April*), 305–318. Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

OVERCONFIDENT: DO YOU PUT YOUR MONEY ON IT?*

Erik Hoelzl and Aldo Rustichini

A group exhibits overconfidence if significantly more than half the group members declare to be better than the median in some characteristic. Overconfidence was found in verbal reports for a variety of characteristics and settings but was less often studied for choice behaviour. In an experiment we tested how perceived relative skill influences verbal and choice behaviour. Treatments varied task difficulty and payment. Choice behaviour changes from overconfidence to underconfidence when the task changes from easy and familiar to non-familiar. This effect is significant when monetary payments are at stake and weak when they are not.

People may be overconfident in many different ways: they may overestimate their abilities; they may perceive themselves more favourably than others perceive them; or finally, they may perceive themselves more favourably than they perceive others. An important example of the latter bias is the observation that a majority of people estimates their skills or abilities to be better than the median. This is the meaning of *overconfidence* that we adopt here. Clearly a widespread overconfidence would have important consequences in economic behaviour. In fact, the literature in economics and finance on theories, models, implications and practical effects of overconfidence is growing to be large and influential: see for instance Benabou and Tirole (2002, 2003), Daniel *et al.* (1998), Gervais and Odean (1998), and Weinberg (2002). A detailed summary of the micro foundation of behavioural finance by De Bondt and Thaler (1995), significantly states that 'perhaps the most robust finding in the psychology of judgment is that people are overconfident' (p. 389).

There is indeed a large and consistent body of evidence in psychology and social psychology supporting this statement. The effect has sometimes been labelled *better-than-average* and has been demonstrated on trait ratings (Alicke, 1985; Dunning *et al.*, 1989) as well as on behaviour ratings (Messick *et al.*, 1985). Similarly, *unrealistic optimism* seems a robust behavioural feature when people have to evaluate their own risk to become victims of unfortunate life-events, compared to the average population (Hoorens and Buunk, 1993; Perloff and Fetzer, 1986; Weinstein, 1980). An even more paradoxical result is the one provided by Klar and Giladi (1997). Their experiments show an 'everybody's better than their group average' effect, where subjects judge each member of their group to be better than most other people in the group, even when no helpful information was provided about him.

The literature in economics has typically taken this evidence as given, and explored its implications. A critical evaluation of the factual and experimental

* We thank Elisabeth Himmer for her help in data collection, and two anonymous reviewers for constructive comments. Part of this paper was prepared while the first author was visiting scholar at Carnegie Mellon University, funded by the Erwin-Schroedinger-Fellowship J2187-G04 from the Austrian Science Fund (FWF). Part of this research was supported by the NSF grant NSF/SES-0136556.

evidence supporting the main hypothesis may be useful. While the results are generally significant, this evidence has some critical features that leave the interpretation of the results in question.

0.1. *The Evidence and its Limitations*

Most of the evidence is based on verbal statements of subjective estimates, not on the observation of choices among alternatives. For instance, in the classical study of Svenson (1981) subjects were asked to rate the safety of their driving, on a percentile scale with 10% interval increments. Svenson (1981) found that 77% of the subjects in one study (Swedish subjects), and 88% in another (US subjects), rated themselves better than the median. But subjects who are simply asked to evaluate their ranking have no incentive to be accurate. The effect of monetary or more generally extrinsic compensations on the accuracy of evaluations and the effort provided by subjects in experimental situations are still debated: see the reviews of Smith and Walker (1993) and Camerer and Hogarth (1999) on the topic. Significant differences in behaviour when rewards are given or not are possible (examples are in Gneezy and Rustichini, 2000*a*,*b*).

The problem of possible lack of control on the subjects is compounded by the potential ambiguity of the questions. For instance, 'safe driver', may be interpreted by different people in very different ways. Driving safely may mean *driving very slowly* to me and *take bends at high speed without going into the ditch* to you. The hypothesis that some of the inconsistencies found in the literature derive from different definitions that people have in mind has been tested by Dunning *et al.* (1989), who show some support for this explanation.

An additional ambiguity in interpretation of the answers may arise because subjects typically have an imprecise or confused idea of concepts like percentile and median, or even average. They may fail to distinguish clearly between average and median, and fail to appreciate the difference between the two concepts for distributions that are not symmetric. An additional ambiguity may surround the idea of the population used for comparison. In many studies, subjects are asked to compare themselves to an average peer, so that people are free to choose a comparison in a lower rank, like a person who is known to be worse off, or (in the studies on optimism) at higher risk. Perloff and Fetzer (1986) and Hoorens and Buunk (1993) for instance showed that the bias was reduced when the closest friend was used as specific comparison target. Weinstein (1980) showed that heightening the awareness that others were in a similar position decreased unrealistic optimism. People might also avoid specific comparisons, but simply apply a self-serving heuristic. Alicke *et al.* (1995) argued that the reality constraints that are imposed by more direct comparisons diminish the better-than-average effect. In their experiments, they showed that by individuating the target and providing personal contact the magnitude of the effect decreased.

Finally one has to keep in mind that the evidence on overconfidence usually refers only to skills in activities for which subjects already have some familiarity. A

privileged role has gone for instance to driving, which actually spurred the early research on the topic (see for instance Preston and Harris 1965; Näätänen and Summala, 1975; Svenson, 1978; Slovic *et al.*, 1978). We know less of how people evaluate their relative skill in non-familiar tasks. An important exception is Griffin and Tversky (1992), who found that experienced subjects have a higher degree of overconfidence.

In summary this evidence leaves open the possibility that if we test the degree of confidence in the relative skill with choices rather than verbal statements, and with tasks of different degrees of difficulty, then the overconfidence may turn into under-confidence. Since some of the most interesting applications in economics and finance concern behaviour where these two features are present, this seems to be a crucial test.

0.2. *Our Experimental Design*

We designed an experiment aiming at a precise test of the overconfidence hypothesis trying to avoid the ambiguities we mentioned.

In our experiment, subjects have to choose among options, not utter statements. The revelation of the relative position that a subject thinks to have with respect to others is deduced from the choice itself. Subjects are free to hold their opinion on what is important in solving a task; what is relevant in their choice is the estimated relative ability to solve it. The subjects relevant for the comparisons are unambiguously the other subjects in the room at the time of the experiment. Subjects are chosen from a homogeneous and familiar environment (their university) so that some idea about their ability is possible. The difficulty of the task can be manipulated: In our experiments, two types of task are used, one more difficult than the other.

In Section 1 we present the design of the experiment. The results of the experiment are reported and analysed in Section 2. We draw our conclusions in Section 3 and, in the Appendix available on the JOURNAL's website www.res.org.uk, we provide background information and further tests.

1. Experimental Design

1.0.1. *Overview*

The design of the study was a 2 (payment: money vs. no money) $\times$ 2 (task difficulty: easy vs. difficult) between-subjects design. In the money condition, participants could win 150 ATS (Austrian Schilling: 150 ATS were worth approximately 10 USD at the time). In the no money condition, they were asked to imagine that they could win the same amount. Before performing the task, subjects had to choose by a majority vote the procedure to determine their payment. In the easy condition the task for the subjects was easier than in the difficult one.

1.0.2. *Subjects*

A total of 134 subjects, 88 women and 46 men, participated in the study. On average, subjects were 23 years old. Six participants stated a profession, four were still going to school, and the remaining participants were students. Of those 117

participants, 66 studied psychology, while the others studied various subjects such as history, languages, medicine, or sociology.

1.0.3. *Task*

Task difficulty was manipulated by choosing simple and difficult items from a test of knowledge of vocabulary. The test is called LEWITE, and is presented in Wagner (1999). It is based on probabilistic test theory and conforms to the Rasch model (Fischer and Molenaar, 1995). The item difficulty parameters derived under this condition are independent from the sample of persons who worked on them. Furthermore, it is possible to compare the ability of persons who worked on different item sets. This test therefore provides a high-quality measure of word-power.

For the selection 126 items were used, with different degrees of difficulty. For the easy task, 22 items were chosen from the lower third; for the difficult task, 22 items were chosen from the upper third of the item pool. Two items in the upper and lower third of each condition were taken as examples, leaving 20 items for each test.

Participants were told that the test would be on their vocabulary and that they would have to explain easy (or difficult) words by completing sentences with gaps. To do so, they had 7 to 9 alternative options for each word, from which they had to choose two to complete the sentence. Subjects in the easy task saw the examples *anode* and *honorarium*; those in the difficult task saw the examples *apse* and *Baleares* (*Webster's New World Dictionary* defines *apse* as 'a semicircular or polygonal projection of a building, especially one at the east end of a church with a domed or vaulted roof' and *Baleares* as 'a Spanish province comprising the Balearic Islands, the latter located in the Mediterranean off the East coast of Spain'). Participants were told that the test would consist of 20 easy (or difficult, in the other treatment) words to explain in this manner.

1.0.4. *Procedure*

Subjects were asked for participation in an experiment about decision making in classrooms and the local cafeteria. No course credits were given for participation. When a sufficiently large group was gathered (only one group has seven subjects, all the others have at least nine), subjects were led into the laboratory and they sat at separate tables.

The questionnaire set consisted of seven pages which were distributed to the subjects in sequence. They were first informed, in the two different treatments, that they could win 150 ATS or that they should imagine winning this amount. They were then told that they would have to vote between two conditions: 'performance test' and 'lottery'. In the condition 'test' a subject would win if his result were in the upper half of the results of all participants. In the 'lottery' condition he would win with 50% probability, with the outcome determined by an individual toss of a die. It was clear that the vote determining the condition was going to be confidential, and the outcome determined by majority vote. A detailed description of the instructions is provided in the Appendix (A.1).

The test and the lottery were both going to be performed independently of the outcome of the vote. The test result was revealed to the subject, but otherwise kept confidential. The confidentiality of the vote, the individual feedback of test results

and the announcement of the average group results alone were chosen to avoid social prestige effects. In order to control for effects of effort avoidance, both test and lottery had to be completed regardless of the voting result.

The voting procedure was clarified by an example. Four more examples were given, which participants had to complete on their own, and correct solutions were provided to ensure proper understanding of the voting procedure. Then the test itself was explained and two sample items were provided. After reviewing the rules for winning in the two conditions 'test' and 'lottery', subjects had to vote for one of the conditions.

After the vote, subjects answered a few questions on the firmness of their decision and gave performance predictions for themselves and the group. Then the real test was performed. The individual test result and the result of the die toss were recorded. Participants were asked about their satisfaction with both results, and how sure they were about their decision now and how easily they could change it. Furthermore, they had to rate the difficulty of the test in retrospect and to estimate the average score of the group. A more detailed description of the time sequence is given in the Appendix (A.2).

1.1. *Discussion of the Design*

The voting mechanism to decide between test and lottery may seem roundabout. In the following section, the rationale for this approach will be discussed in detail, and compared to several possible alternatives.

1.1.1. *Different options*

A possible alternative design of the experiment is to ask each person to choose between the test or the lottery as the way to decide his payment, and to consider only the score of those who voted for the test. This choice has a clear and serious problem. At equilibrium, subjects who consider themselves at the bottom of the scale of skill in answering the questions will probably vote for the lottery. So the next subject up is now the bottom of the scale and he will vote for the lottery. This process will continue until at equilibrium only the top subject chooses the test. The precise characterisation of the equilibrium depends on a precise specification of the belief subjects have on the distribution of skills in the population. But it is clear that an individual choice makes the choice of test potentially useless as a measure of confidence in one's relative skill.

Another alternative is to proceed the way we just described but considering the performance of all subjects, independently of their choice between test and lottery. This solution has the inconvenience that subjects who chose the lottery are likely to put very little effort (or no effort at all) in the task, since they have no incentive to do so.

1.1.2. *The vote procedure*

The solution of having the subjects vote on the two options seems to give a satisfactory solution to the problems inherent in the two alternative solutions described above. If a subject attaches a probability larger than 50% to the event that he will

be in the top half of the subjects in the test, then voting for the test is for him a weakly dominant strategy. At least this is true if we abstract for the moment from risk and ambiguity aversion: but on this point we return in the conclusions. So the voting mechanism is one way to ensure a truthful revelation of how subjects rate their relative skill.

1.1.3. *The equilibrium of the game*

The solution of the incentive problems through voting has a price. Consider first a subject who is comparing the merits of voting for the test or for the lottery. He has some idea of the difficulty of the task, some idea of his skills in it, and some idea of the skills of the other participants in the experiment. The skill may be described for convenience (and this assumption is probably close to the real evaluation process of the individual) by a single parameter. If our subject thinks that all the others follow the strategy of voting for the test when and only when their skill parameter is higher than the median, then it is best for him to follow that strategy too. So this strategy profile is indeed an equilibrium. However, it is not the only one. The reason for this is discussed in detail in the Appendix (A.3).

Ignoring the details for the moment, we can see that the intuition for the result is clear. The vote of a player only matters when he is pivotal, that is when the equilibrium cut-off point is the median type in the set of all the other players. But in this event if our player's type is higher than the cut-off point then he is also higher than the median type, and so he should vote for the test. In other words, the intuition that he should vote for the test when his type is higher than the median is correct. However the relevant median value is determined by the voting, and can be anywhere the equilibrium sets it.

Of course, the equilibrium where all players vote for the test when they think they are above the median has a special status, of focal point. But the multiplicity of equilibria leaves us with the additional task of checking, in the data of the experiment, which equilibrium subjects play. This is done in Section 2.3 below.

2. Analysis of the Data

2.1. *Subjects' Predictions*

The predicted performance (that is, the number of correct answers) that each subject gave for himself before the test was 12.51, less than the value of 13.37 predicted for the group (see Table 1); so subjects predict for themselves a lower performance than for the group by a margin of approximately 5%.

Table 1

Subjects' Performance Predictions, Before the Test

Variable	Mean	st.d error	(95% conf. Interval)
Predicted own performance	12.51	0.30	(11.92, 13.11)
Predicted group performance	13.37	0.24	(12.90, 13.83)

Note: $n = 134$.

The average number of correct answers for each person was 11.83, lower than the predicted own performance (see Table 2). The error, however, was only 5%. The prediction appears to be accurate, even remarkably so, if one considers that subjects were typically new to the task and had no previous quantitative estimate available on their accuracy. Observing a performance lower than expected did not lead them to revise the estimate on the performance of the group downwards (see Table 2). The average estimated value for the group after the test was 13.43, which is statistically the same as before the test.

These qualitative features of the behaviour for the entire sample of subjects are by and large unchanged if we condition on the different treatments. The predictions made before the test are reported in Tables 3 and 4. The ratio between expected own performance and the expected group performance changes with the difficulty of the task. However, the ratio changes also with the payment treatment, which already indicates that subjects are less willing to assert their skills when money is at stake.

The analysis by treatment also confirms that subjects fail to adjust the prediction on the group performance downwards after observing that their performance on

Table 2

Subjects' Performance and Estimated Group Performance, After the Test

Variable	Mean	st.d error	(95% conf. interval)
Actual own performance	11.83	0.45	(10.94, 12.72)
Estimated group performance	13.43	0.30	(12.84, 14.03)

Table 3

Subjects' Performance Predictions and Ratio, by Payment

	No money treatment		Money treatment	
Variable	Mean	st.d error	Mean	st.d error
Predicted own performance	13.25	0.40	11.83	0.43
Predicted group performance	13.66	0.35	13.09	0.32
Better	0.99	0.03	0.91	0.03

Note: The variable *better* is the individual ratio between *predicted own performance* and *predicted group performance*.

Table 4

Subjects' Performance Predictions and Ratio, by Difficulty

	Easy treatment		Difficult treatment	
Variable	Mean	st.d error	Mean	st.d error
Predicted own performance	13.80	0.35	11.06	0.44
Predicted group performance	13.92	0.28	12.75	0.38
Better	1.00	0.02	0.88	0.03

Table 5

Subjects' Performance and Estimated Group Performance After the Test, by Payment

	No money treatment		Money treatment	
Variable	Mean	st.d error	Mean	st.d error
Actual own performance	12.34	0.65	11.35	0.63
Estimated group performance	13.80	0.40	13.09	0.45

Table 6

Subjects' Performance and Estimated Group Performance After the Test, by Difficulty

	Easy treatment		Difficult treatment	
Variable	Mean	st.d error	Mean	st.d error
Actual own performance	16.14	0.28	6.97	0.33
Estimated group performance	16.07	0.22	10.46	0.29

the test was worse than expected (see Tables 5 and 6). Only in the case of the difficult treatment does the predicted value of group performance fall after the test. The prediction on own performance had an average of 11.06; the prediction on group performance had an average of 12.75 before the test and an average of 10.46 after the test.

In Tables 14 and 15 in the Appendix we report the summary statistics for the answers of the subjects to the questionnaire. The challenge of a difficult test does not increase the importance that subjects say a good performance has for them. The variable *important* describes the importance (on a scale from 1 to 7 = very important) of a good result in the test. This variable has 3.24 as overall average, 3.23 in the easy treatment and 3.25 in the difficult treatment.

In view of the fact that the real performance is lower than the predicted one, the finding that subjects are as sure about their decision in the vote after (with an average of 5.33) than they were before (average 5.16) may seem puzzling. But we have already noted how the prediction was after all reasonably accurate. In Table 16 in the Appendix we present the regression of the number of correct answers against the predicted number of correct answers. The coefficient is positive and significant, but the low R-squared shows that there were possibly large individual errors.

2.2. *The Choice of Vote*

Across the different treatments, the pattern of verbal statements about one's expected performance is similar to the pattern of choice behaviour. However, there are some inconsistencies between the verbal statements and the choice. Table 7 summarises the frequency of statements in which the expected own performance is larger than the expected group performance. In other words, this is

Table 7

Frequency of Predicted Own Performance Larger than Predicted Group Performance, by Treatments

	No money treatment			Money treatment		
	Mean	st.d error	(95% conf. interval)	Mean	st.d error	(95% conf. interval)
Easy treatment	0.47	0.08	(0.31, 0.64)	0.30	0.08	(0.14, 0.47)
Difficult treatment	0.33	0.09	(0.14, 0.52)	0.19	0.07	(0.06, 0.33)

Note: Easy, no money $n = 38$; Easy, money $n = 33$; Difficult, no money $n = 27$; Difficult, money $n = 36$.

the average value of a variable which is one if the expected own performance is larger than the expected group performance, and zero otherwise, and is therefore comparable to the vote. The highest value is in the easy, no money treatment and the lowest in the difficult, money treatment. Over all the treatments the value is 0.33, indicating that many participants predict a lower performance for themselves than for the group.

On the other hand, the overall vote is of 55% in favour of the test condition. The average vote (where *vote* equal to 1 if the subject votes for the test, and 0 if he votes for the lottery) for each of the four different treatments is reported in Table 8.

It is clear that going from the easy to the difficult treatment and from the no money to the money treatment discourages the vote for the test. The average vote for the test is 0.63 and 0.64 in the two easy treatments (without or with money), with no significant difference between the two. The average falls to 0.56 in the difficult, no money treatment, and below half, to 0.39, in the difficult, money treatment. The only case in which the *lottery* wins by majority voting is the last, that is the difficult, money treatment. The confidence intervals suggest that the difference between the latter treatment and the two easy treatments is significant.

The non-parametric Wilcoxon-Mann-Whitney test confirms this conclusion. In addition, the test shows that the difference in significance across the different treatments is due to the change from easy to difficult treatment. In Table 9 we present the results of the pairwise comparisons; the first column compares the observations in the no money treatment with those in the money treatment, while the second does the same for the easy and difficult treatments. The first comparison is not significant, while the second is.

A more systematic comparison, across all pairs of different treatments, is provided in Table 10. In the first column (marked 'easy') we compare the distribution

Table 8

Average Vote for Test, by Treatments

	No money treatment			Money treatment		
	Mean	st.d error	(95% conf. interval)	Mean	st.d error	(95% conf. interval)
Easy treatment	0.63	0.08	(0.47, 0.79)	0.64	0.09	(0.46, 0.81)
Difficult treatment	0.56	0.10	(0.36, 0.76)	0.39	0.08	(0.22, 0.56)

Table 9

Wilcoxon-Mann-Whitney Non-parametric Tests of Vote for Test, by Difficulty and Payment

	No Money : Money	Easy : Difficult
z	1.08	2.01
p-value	0.28	0.04

Note: The H_0 is equal distributions of vote between different treatments.

Table 10

Wilcoxon-Mann-Whitney Non-parametric Tests of Vote for Test, by Treatments

	No Money : Money		Easy : Difficult	
	Easy	Difficult	No money	Money
z	−0.04	1.30	0.61	2.04
p-value	0.97	0.19	0.54	0.04

Note: The first two columns test for equal distributions of vote between the No Money and Money treatments, for the two levels of difficulty. The second two columns test for equal distributions of vote between the Easy and Difficult treatments, for the two levels of payment.

of the vote for the test in the no money *versus* the money treatment, for the subjects in the easy treatment, while in the second we make the same comparison for subjects in the difficult treatment. In the third column (marked 'no money') we compare the distribution of the vote for the test in the easy *versus* the difficult treatment for subjects in the no money treatment, and in the fourth the same comparison for subjects in the money treatment is made.

The difference is significant in the case of the money treatment, when the comparison is made between the easy and difficult treatment. The other differences are not significant. So difficulty alone does not change the behaviour. In the no-money condition, the vote is not significantly different in the easy and difficult treatment. Only when payment is offered the vote is significantly different in the difficulty levels.

2.3. *Choice of Equilibrium*

We have seen in Section 1.3, *The Equilibrium of the Game*, that the equilibrium where people vote for the test when their own perceived skill is larger than the median is one of many equilibria. This equilibrium is however the most natural. In this Section we check that this is indeed the equilibrium that is being played.

The logit regression of the vote over *better* (which is the ratio between *predicted own performance* and *predicted group performance*) gives a positive relationship between the probability of voting for the test and the variable *better* (see Table 11).

According to the results in Table 11, the probability of voting for the test in the aggregate sample has the form:

Table 11

Logit Regression of Vote for Test Over Ratio of Performance Predictions

	pseudo- R^2	χ^2	Prob > χ^2
	0.0603	11.11	0.0009
Vote	Coefficient	*z*	*P* > *z*
Better	2.5468 (0.8190)	3.11	0.002
Constant	−2.1868 (0.7896)	−2.77	0.006

Note: The variable *better* is the individual ratio between *predicted own performance* and *predicted group performance*

$$\mathrm{P}(vote = 1) = \frac{\exp[2.5468(better - 0.8568)]}{\{1 + \exp[2.5468(better - 0.8568)]\}}$$

if we factor out the coefficient of the variable *better*. So the inflection of the logistic curve is at the value 0.8568 of the variable *better*, which is close to 1.

The logit regression with the two variables *predicted own performance* and *predicted group performance* as predictors also gives significant coefficients, of the expected sign (positive for *predicted own performance* and negative for *predicted group performance*; see Table 12).

The significance of the relationship changes greatly, however, as we consider specifically different treatments. It is the *least* significant for the difficult, money treatment (Table 17 in the Appendix). Although the coefficients are of the expected sign, they are not significant.

2.4. *Summary of the Results*

We can summarise our main findings as follows:

First, there is an important and significant difference in the behaviour when the task is perceived to be difficult rather than easy. Going from the easy to the difficult treatment the average vote shifts from a majority going to the test (63%) in the easy case to a significantly smaller fraction going to the test (56% in the no money, and 39% in the money, with an average across subjects of 46% over these two treatments) in the difficult case.

Table 12

Logit Regression of Vote for Test Over Performance Predictions

	pseudo-R^2	χ^2	Prob > χ^2
	0.0708	13.04	0.0015
Vote	Coefficient	*z*	*P* > *z*
Predicted group performance	0.2186 (0.0666)	3.28	0.001
Predicted own performance	−0.2036 (0.0848)	−2.40	0.016
Constant	0.2003 (0.9445)	0.21	0.832

Second, the difficulty of the task is mainly of importance when money is at stake. When payment is offered and the task is difficult, the proportion of subjects voting for the test is significantly lower than when the task is easy. This conclusion seems to question the applicability of findings from related studies to real-world tasks in which both the stakes are higher and the problem at hand is more complex than in the laboratory.

The dependence of the behaviour on the difficulty of the test is important if subjects actually do perceive the harder test as being in fact harder. The regression of the perceived difficulty (*difficult test*) against the actual difficulty shows that the relationship is indeed strong (the coefficient is 0.97) and significant ($t = 5.13$, p-value $= 0.0001$). This is not surprising, given that in addition to the examples the instructions clearly label the easy and difficult tasks as such. The test we have just described confirms that the information provided by the examples and the statement in the instructions is effective in shaping the expectations. However, within each treatment, the relation between statements of perceived difficulty and voting behaviour is less clear. Surprisingly, a logit regression of the vote over *difficult test* yields a positive and significant coefficient (1.22, $z = 2.25$, p-value $=$ 0.024) in the difficult, money condition, and negative, non-significant coefficients in the other conditions. Although these findings may not be robust, they seem to leave some doubt about the usefulness of verbal statements for determining the degree of confidence.

3. Conclusions

The aim of the paper was to provide a reliable measure of the degree of confidence that subjects have in their skill, relative to others and to test it experimentally. The measure we suggested is the fraction of subjects voting in favour of a reward depending on performance in a test rather than a reward depending on pure chance. Since only half of the subjects will win if the test decides the winner, any excess over a half of the subjects voting for the test indicates an erroneous evaluation of their own relative skills.

This measure seems effective but it has an inherent bias. Subjects in the experiment are facing the choice between a lottery with objective uncertainty, and an ambiguous choice. The uncertainty that they face when the payment is decided by the test is very similar to the one faced by subjects in the choice presented by Ellsberg to his hypothetical subjects. The distribution of skill and talent of the opponents in the test is unknown, as much as the distribution of the balls of different colour in Ellsberg's 'experiment'. So our suggested measure of confidence is likely to underestimate the subjective perception that subjects have of their skill relative to other players, due to ambiguity aversion. In addition, the voting game when interpreted strictly as a game between rational players has, as is typical with any voting situation, many different equilibria.

We provided an estimate of this measure in a laboratory experiment, based on a simple word-knowledge task. Two different treatments (easy *versus* difficult task, and payment *versus* no payment) were used. The main results seem to indicate that the hypothesis of a universal bias to overconfidence is not supported.

To start with, there is some difference between how subjects act and how they perceive (or at least state to perceive) their relative ability. For instance, the average fraction of subjects who expect their performance to be higher than the group performance is only 0.33, while the average vote for the test is 0.55.

More importantly, there is a sharp difference between the vote, hence according to our measure the degree of confidence, in the easy and difficult treatment. This difference is particularly strong when payment is offered, even if the amount at stake was rather modest. When the effects of difficulty and payment are combined, subjects are clearly underconfident.

Two interpretations of this result are possible. The first interpretation is that subjects confuse 'being good' with 'being better'. Subjects facing an easy test anticipate being able to solve a large number of questions, while they expect to solve a smaller number when facing a difficult test. They fail to make the same adjustment for the rest of the population, and in particular for the sample of their opponents. So their predicted estimated relative ability is greater when they have an easy task. This interpretation is in line with empirical findings by Klar and Giladi (1999), Kruger (1999) and Eiser *et al.* (2001), suggesting that in comparative judgments, people focus primarily on self-assessments and less on assessments of others. In a similar vein, the results by Camerer and Lovallo (1999) indicate that people insufficiently adjust for characteristics of their competitors.

A second interpretation is possible in terms of ambiguity aversion. Heath and Tversky (1991) found that people prefer to bet on their own judgment when they are knowledgeable in a domain, and on chance otherwise. Subjects facing a relatively more difficult task also believe that their knowledge of the distribution of ability for that specific task is smaller than the one they have in the case of an easy task. If I know less about quantum physics than I know about television programmes, I am also likely to think that I know less about how much quantum physics is known in the population at large. So the ambiguity subjects are facing with a difficult task is larger, hence their choice shifts in the direction of the lottery that protects them from the additional ambiguity.

University of Vienna
University of Minnesota

Date of receipt of first submission: June 2002
Date of receipt of final typescript: May 2004

A Technical Appendix is available for this paper at: http://www.res.org.uk/economic/ta/tahome.asp
The dataset is available for this paper at: http://www.res.org.uk

References

Alicke, M. D. (1985). 'Global self-evaluation as determined by the desirability and controllability of trait adjectives', *Journal of Personality and Social Psychology*, vol. 49(6) (December), pp. 1621–30.

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J. and Vredenburg, D. S. (1995). 'Personal contact, individuation, and the better-than-average-effect', *Journal of Personality and Social Psychology*, vol. 68(5) (May), pp. 804–25.

Benabou, R. and Tirole, J. (2002). 'Self-confidence and personal motivation', *Quarterly Journal of Economics*, vol. 117(3) (August), pp. 871–915.
Benabou, R. and Tirole, J. (2003). 'Intrinsic and extrinsic motivation', *Review of Economic Studies*, vol. 70(3) (July), pp. 489–520.
Camerer, C. F. and Hogarth, R. M. (1999). 'The effects of financial incentives in experiments: a review and capital-labor production framework', *Journal of Risk and Uncertainty*, vol. 19(1–3) (December), pp. 7–42.
Camerer, C. and Lovallo, D. (1999). 'Overconfidence and excess entry: an experimental approach', *American Economic Review*, vol. 89(1) (March), pp. 306–18.
Daniel, K., Hirshleifer, D. and Subrahmanyam, A. (1998). 'Investor psychology and security market under- and overreactions', *Journal of Finance*, vol. 53(6) (December), pp. 1839–85.
De Bondt, W. and Thaler, R. H. (1995). 'Financial decision-making in markets and firms: a behavioral perspective', in (R. A. Jarrow, V. Maksimovic and W. T. Ziemba, eds), *Finance, Handbooks in Operations Research and Management Science*, vol. 9, pp. 385–410. Amsterdam: North Holland.
Dunning, D., Meyerowitz, J.A. and Holzberg, A.D. (1989). 'Ambiguity and self-evaluation: the role of idiosyncratic trait definitions in self-appraisals of ability', *Journal of Personality and Social Psychology*, vol. 57(6) (December), pp. 1082–90.
Eiser, J. R., Pahl, S. and Prins, Y. R. A. (2001). 'Optimism, pessimism, and the direction of self-other comparisons', *Journal of Experimental Social Psychology*, vol. 37(1) (January), pp. 77–84.
Fischer, G. H. and Molenaar, I. W. (1995). *Rasch Models: Foundations, Recent Developments, and Applications*, New York: Springer.
Gervais, S. and Odean, T. (1998). 'Learning to be overconfident', *Review of Financial Studies*, vol. 14(1) (Spring), pp. 1–27.
Gneezy, U. and Rustichini, A. (2000*a*). 'A fine is a price', *Journal of Legal Studies*, vol. 29(1) (Part 1 January), pp. 1–18.
Gneezy, U. and Rustichini, A. (2000*b*). 'Pay enough or don't pay at all', *Quarterly Journal of Economics*, vol. 115(3) (August), pp. 791–810.
Griffin, D. and Tversky, A. (1992). 'The weighing of evidence and the determinants of overconfidence', *Cognitive Psychology*, vol. 24(3) (July), pp. 411–35.
Heath, C. and Tversky, A. (1991). 'Preference and belief: ambiguity and competence in choice under uncertainty', *Journal of Risk and Uncertainty*, vol. 4(1) (January), pp. 5–28.
Hoorens, V. and Buunk, B. P. (1993). 'Social comparison of health risks: locus of control, the person-positivity bias, and unrealistic optimism', *Journal of Applied Social Psychology*, vol. 23(4) (February), pp. 291–302.
Klar, Y. and Giladi, E. E. (1997). 'No one in my group can be below the group's average: a robust positivity bias in favor of anonymous peers', *Journal of Personality and Social Psychology*, vol. 73(5) (November), pp. 885–901.
Klar, Y. and Giladi, E. E. (1999). 'Are most people happier than their peers, or are they just happy?', *Personality and Social Psychology Bulletin*, vol. 25(5) (May), pp. 585–94.
Kruger, J. (1999). 'Lake Wobegon be gone! The 'below-average effect' and the egocentric nature of comparative ability judgments', *Journal of Personality and Social Psychology*, vol. 77(2) (August), pp. 221–32.
Messick, D. M., Bloom, S., Boldizar, J. P. and Samuelson, C. D. (1985). 'Why we are fairer than others', *Journal of Experimental Social Psychology*, vol. 21(5) (September), pp. 480–500.
Näätänen, R. and Summala, H. (1975). *Road-user Behavior and Traffic Accidents*, Amsterdam: North-Holland.
Perloff, L. S. and Fetzer, B. K. (1986). 'Self-other judgments and perceived vulnerability to victimization', *Journal of Personality and Social Psychology*, vol. 50(3) (March), pp. 502–10.
Preston, C. E. and Harris, S. (1965). 'Psychology of drivers in traffic accidents', *Journal of Applied Psychology*, vol. 49(4), pp. 284–8.
Slovic, P. B., Fischhoff, B. and Lichtenstein, S. (1978). 'Accident probabilities and seat belt usage: a psychological perspective', *Accident Analysis and Prevention*, vol. 10(4) (December), pp. 281–5.
Smith, V. L. and Walker, J. M. (1993). 'Monetary rewards and decision cost in experimental economics', *Economic Inquiry*, vol. 31(2) (April), pp. 245–61.
Svenson, O. (1978). 'Risks of road transportation in psychological perspective', *Accident Analysis and Prevention*, vol. 10(4) (December), pp. 267–80.
Svenson, O. (1981). 'Are we all less risky and more skillful than our fellow drivers?', *Acta Psychologica*, vol. 47(2) (February), pp. 143–8.
Wagner, M. (1999). 'Lexikon-Wissen-Test (LEWITE)', unpublished dissertation, University of Vienna.
Weinberg, B. (2002). 'A model of overconfidence', Ohio State University Discussion Paper.
Weinstein, N. D. (1980). 'Unrealistic optimism about future life events', *Journal of Personality and Social Psychology*, vol. 39(5) (November), pp. 806–20.