Past success and present overconfidence

Novarese, Marco

Center for Cognitive Economics - Università Amedeo Avogadro

March 2009

Past success and present overconfidence

Marco Novarese
Centre for Cognitive Economics, University Amedeo Avogadro
marco@novarese.org

February 2009

Preliminary version

Comments are welcome

Abstract

According to a wide literature persons are not able to evaluate their own skills and knowledge, but the discussion on the nature, extension and determinants of this phenomenon is still open. This paper aims at proposing new empirical evidence on overconfidence and its determinants, trying to find out the possible effect of past performance on present optimism. I test my students' calibration and confidence in predicting their future results, comparing their expectations and real grades. My analysis allows showing the existence of overconfidence, its reduction in two following tests, and its non linear relation with students' capacities. Besides, I focus my attention on the effect of the grade my students got at the end of high school. This is used a proxy of their past experience and habit to get good or bad grades. Past success determined overconfidence. This idea is connected to the literature on heuristics and rule based perception.

Keywords

overconfidence, expectations, learning, inertia, rule based behavior, economic education

Jel codes

D83

Past success and present overconfidence

1. Introduction

This paper aims at proposing new empirical evidence on overconfidence, focusing on a new possible determinant. A wide literature evidences that people are not able to evaluate their own skills and knowledge and often tend to be excessively optimist, but the discussion on the nature, extension and determinants of this phenomenon is still open. My analysis is based on a kind of natural experiment. I test, in fact, my students' calibration and confidence in predicting their future results, comparing their expectations and real grades. This data allows testing for the existence of overconfidence and for its relation with students' capacities. Besides, I focus my attention on the effect of the grade they got at the end of high school, used a proxy of their past experience and habit to get good or bad grades and therefore of their previous performances in study. So I can study if past success influenced my students' overconfidence and capacity to calibrate their expectations. This idea is inspired from the literature on heuristics and rule based perception and allows linking the analysis of overconfidence with a wider literature on rationality.
The first part of the paper recalls some results of the literature on overconfidence and its determinants, focusing on the effect of knowledge and capacity and rule following behavior (§2.1). I then presents some results on student's overconfidence (§2.2). Then I present my data (§3) and propose some hypothesis (§4), which are then tested (§5) and discussed (§6).

2. Background

2.1. Overconfidence and experiments

In many cases, people seem to be excessively confident in their abilities and optimistic about their chance of success. Van Den Steen (2004) recalls a wide series of different results and gives a rational justification of what he calls over-optimism, a general term which includes many kinds of biases. Overconfidence is one of the way in which over-optimism takes place, and usually indicates the tendency to over-estimate the correctness of one's answers. Confidence in one's knowledge is usually assessed asking persons a series of questions and at the same time requiring them to state their confidence on the correctness of their previous replies. As noted by Klayman et al (1999), in a first phase, researchers agreed on the idea that persons are usually overconfident and systematically over-estimated their chances of giving correct answers. When stating a level of confidence equal to, say, 90%, persons were seen as able to answer correctly at a very lower percentage (like 75%). Some human cognitive and motivational attitudes were and still are called for explaining these results. When trying to answer a question on which they are not sure, persons search in their memory for cues. The mechanism of associative memory could make it easier remembering information which confirms one's own initial ideas. The first impression could, therefore, appears to be more founded than it is, with a following increase in (over)confidence (Rabin and Schrag, 1999; Russo and Schoemaker, 1992). Subjects, besides, like thinking to be smart and to be correct and therefore could tend to over-estimate their chance of correctness.
Lichtenstein and Fischhoff (1977), yet, show a less linear relation in a similar experiment. Subjects were, as usual, requested to state a level of confidence for their answers to a cultural test. Their analysis shows how those who knew less are overconfident, but those who knew more could be even underconfident. They found no individual differences related to intelligence or expertise. So, *those who know more (on a specific question) doesn't really know more on how much they know*. Gigerenzer (1991) confirms and explains their findings. In his view, subjects take decision using simple but smart heuristics. Many of the tests used in these kind of experiments, for example, could be faced using a specific rule, defined the *recognition heuristics*. Frequently participants to these experiments are required, for example, to indicate which city, between two proposed, has more inhabitants. They should also state a probability of correctness for each of their answers. Perfect calibrated persons should, as an example, reply correctly to half of the questions for which

their confidence is stated as equal to 50%. Systematic differences between the two values can indicate overconfidence or underconfidence.

When applying the *recognition heuristic*, persons identify the most known city as the biggest one. So, a city which has a soccer team playing in the main football league could be more recognizable than a city which have no such team and, therefore, more probably identified as the biggest. Gigerenzer (1991) shows that this simple procedure forecasted well real individual responses. He also proves that such a simple rule is actually smart, as it is able to give correct answers in about 91% of the cases, when, for example, two German cities are randomly selected and coupled. Overconfidence can emerge, in some cases, when this rule is applied. Subjects, in fact, can perceive they have a chance of success of about 91% and state it as their mean level of confidence, if required. If experimenters select specific couples of city, instead of using random extractions, this mean level of confidence could, yet, be wrong, and subjects could result overconfident or underconfident. According to Gigerenzer (1991), usually experiments propose just tricky couples of cities, so determining at the same time errors and overconfidence. In Gigerenzer's experiment, on the contrary, overconfidence disappeared as random couples of cities are proposed. If the biggest among the two selected cities is identifiable using the recognition heuristics, subjects are even underconfident, as they refer again to their rule general rate of success. According to Gigerenzer (1991) in most cases, experiments are, therefore, based on biased questions: too difficult ones are often selected.

In these kinds of experiments, subjects can be required to evaluate their performance in two different ways: they can indicate their confidence for each single question, or give an overall rate for the whole test. According to Gigerenzer (1991), the procedure used is relevant, as it can activate the use of different heuristics. The *recognition heuristics* cannot be used in all cases, in fact. If required to evaluate their overall performance (instead of indicating a case by case confidence), subjects could not be able and used to compute a mean value among different percentages[1]. They would probably use another reference point, as their mean rate of success in previous tests perceived as similar. Again, if the set of questions is not randomly selected, subjects would probably evaluate in a bad way their performance.

More in general, take a person who usually experiences a rate of success of, say, 80%, in answering a specific kind of random selected question. When answering a biased set of questions, the rate of success is different. If the questions are selected among the easiest ones (for any reason; in a natural environment question could be easier because our agent is actually facing a different situation), the subject will have a higher rate of success than expected. If selection favours difficult items, the rate will be lower. This is true if subject are not able to evaluate that a specific set of questions is not random. So, according to this idea, those who know more are not necessarily most capable of evaluating their knowledge. Similarly persons used to have a god performance in a given task, could have an optimistic expectation for their performance in a more difficult task they perceive as similar to the previous one.

This idea is coherent with a wide literature on rule following behaviours. A bounded rational agent is not able to deal with all information faced and to solve hard problems (Vanberg, 1993 and Heiner, 1983). Rules allow dealing with a complex environment as they provide a reliable action which can be activated in response to a limited set of information, as the recognition heuristic exemplifies. Rules require relatively stable environments and also the expectations of stability. I cannot apply today the strategy I learned yesterday, if I think that the world is now different. So rule following is based on some kind of intelligent inertia in the expectations (Rizzello, 1999) which can have also an emotional counterpart. In the managerial literature, sometimes we can find the expression "success trap", to indicate how some good practices, developed in a specific environment, can fail in different situations (Levinthal and March, 1993). If agents are not able to recognize situations, a series of successes can be followed by some losses. While stable expectations can be often useful, and even necessary to learn (Novarese, 2009), they could also determine decisional errors and a wrong perception.

---

[1] If all answers can be given using the same heuristics, the mean value can be easily computed.

Because of their nature, beside, rules have to be generic (or they couldn't be applied in different situations). So, almost inevitably, some situations can activate more than one rule. Rules, in fact, have to be domain specific and so it could be normal to combine them. If a decision require different skills, in fact, more than one rule need to be applied, as stated by Hayek (1969). The reality is complex, rules are abstract and specific. While in some occasion there could be just one clear rule, in other occasions, different bits of knowledge need to be applied.

These ideas can help in understanding how students form their expectations on their future grade.

2.2. The overconfidence of students

Overconfidence and over-optimism have also been used to explain student's evaluation of their capacities, their expected grades, and the following performances and satisfaction. Bengtsson et al (2004), as an example, tested overconfidence, analysing if their students responded to the last demand of their written economics test. Such last question was devoted only to students who got at least a given rate in the previous part of the test. Those who spent time and energy in answering a question which gave them no score were overconfident. Males showed this behaviour in a higher degree than females. By construction, in this test, overconfidence was necessarily associated with bad results.

Ochse (2001) asked her students, to state their expected grades before an exam, and then compared it with their real ones. She also required her students to state a level of confidence for the correctness of their same previsions. The most optimistic expectations (i.e. the ones with the highest difference from the real grade), and the highest levels of confidence were associated with worst performances. Ochse suggested the existence of a causal relation going from a high self perception to lower academic achievements.

It's worth noting that forecasting a grade is just an assessment of one's knowledge, quite similar to the indication of the percentage of correct answer in a given test. So, Oches required an evaluation of the same overconfidence (if forecasting the grade is seen the main task, the overconfidence need to be related to such a task).

Grimes (2002) asked his students in Economics to predict their second mid term test score. He found a strong degree of overconfidence: most of the students' expectations were higher than their real scores and very badly calibrated (calibration is measured as the relative difference between expected and real grade). The main determinants of overconfidence resulted to be:

- age and experience (his sample included students of different ages and faculties, older and more mature ones were less overconfident and better calibrated);
- academic abilities; confirming the finding of previous studies, academic and educational achievements proved to be related with less overconfidence and better predictive calibration;
- students who had completed previous economics classes, in high school or college, had higher chances of over-predicting their performance;
- the difference between student's score in the first mid term exam and the average class's score; getting better scores (in the specific course under evaluation) reduced again the probability of overconfidence.

So Grimes found a clear negative correlation between students' capacities (both in his course and in other previous academic courses; a student can have good mean grade, but a bad performance in Economics, so the two kind of indicators can differ) and overconfidence. In his view, overconfidence and bad calibration are caused by limits in meta-cognitive capacities, that are just the capacity to self evaluate one's own knowledge and learning processes. Among cognitive psychologists there is a discussion on the real meaning and significance of meta-cognition. It should be related to the capacity to exert control on what understood, when learning and "how to approach a given learning task, monitoring comprehension, and evaluating progress toward the completion of a task" (Livingston, 1997). So, the meta-cognitive skills would determine more controlled process and, therefore, should, at the same time, increases both the capacity to self evaluate and the performance. Yet, it is possible to imagine that meta-cognition have even a different effect. A better capacity to evaluate what I know, can make me aware of what I don't know and determine underconfidence.

3. The data

My analysis is based on the data gathered on the students who attended my course in Economics, at the Faculty of Law, at the *Università del Piemonte Orientale Amedeo Avogadro*, in the Academic Year 2005-2006. The sample used for the analysis includes all the 57 students who sustained the two midterm written examinations in Economics. All of them were just starting their academic experience, and represent a highly homogeneous sample. My Economic course had two midterm tests. These exams were open only for the persons who attended at least three quarter of the lessons. A week before any of them, students were required to state their expected grade for the test. All the information were collected by a helping student, enrolled in another course, and analyzed only after the exams. In this way I tried to avoid an eventual attempt by my students to over-estimate their own expectations in order to influence my evaluation. A student could have feared that a low expectation would have been seen as a kind of declaration of bad study. So students could have had the incentive to over estimate their expectations. In order to avoid this, I introduced also a reward for the persons whose expectations were correct or wrong by plus or minus half a grade. This incentive was also used as a way to explicitly stimulate a proper self evaluation.
At the beginning of my course I collected also some general information on my students. In particular I asked the grade received at the end of high school (a variable quite relevant here)[2]. Table 1 reports detailed information on the timing of the exams and on the collection of information (some of these data are not used here, but are shown to give a complete picture).

Table1. Timetable: expectations and grade

September 18: lessons start
October 3: Students were asked their grade at the end of the high school (SCOREHIGH)
October 26 : Students declared their expectations for the score in the first exam (EXP1)
November 2: First mid-term test in Economics
November 7: Students were shown the score of their first test in Economics (SCORE1) and were required to state their satisfaction (SODD1)
November 22: New test in economics for the students who didn't pass the first one or were unsatisfied of the score
November 25 Students were shown the score of the new test in Economics
November 30: Students declared their expectations for the score in the second exam (EXP2)
December 7: Second test in Economics
December 10: Students were shown the score of the second test in Economics (SCORE2)

4. Hypothesis and models to be estimated

In respect to the previous recalled analysis, my study presents some differences. My data are collected in a non experimental task, and I required my students to perform an ex-ante evaluation. Yet the studies presented can be anyway a point of departure. In fact, the kind of evaluation requested to my students is similar to the one in standard experiments: in both cases persons are required to state a percentage of correctness. If we apply Gigerenzer (1991)'s idea, there should be no differences between ex ante or ex post evaluations, as both should be based on a same general rate of success of a given rule.
From another perspective, yet, rule following does not necessarily determine an automatic and simple behaviour. Rules can be seen as components of individual knowledge (Vromen, 2004).

2 I required this information at the beginning of my lessons so not to suggest any link with the expected grade for the mid term tests.

Subjects can't be always able to isolate simple contexts where just one simple rule can be activated. As my students' expectations had to be stated a week before the exam, meta-cognition could have played a minor role. In fact, the amount of pages to be studied for both pre-tests was small. High school students are used to study just the few days before their exams. So my students' expectations could have been based also on their self confidence and optimism toward the future. In this perspective, overconfidence could be more a result, than a product of worse knowledge. An aptitude like optimism can be connected to a rule based perception of the world, as it requires constant expectations and perception of the world and some kind of standard response (even something like: "don't worry, everything will be ok") to a situation perceived as similar to a reference one (Bateson, 1973). In doing this evaluation they could have been influenced by their previous experiences at school, a natural similar situation; they could, yet, have associated my exam to some specific course attended and to the relative grade, more than to their general previous experience (a student could have had a low grade in, say, math, and a good overall evaluation and economics could have been perceived as more similar to math than to other disciplines). So, as in the analysis proposed by Grimes (2002) different factors can account for the final results and determine more or less overconfidence and predictive calibration; in other words: the capacity to forecast should be related to different attitudes and levels of skills. I take the estimations proposed by Grimes (2002) as a starting point[3], trying to investigate some new aspects. Overconfidence and calibration are measured here using two variables: PREC and DIFF (followed by a number - 1 or 2 - which indicates if the variable is relative to the first or to the second exam). The variable PREC is used by Grimes and inspired by Lichtenstein and Fischhoff (1977). It is computed as the ratio between the squared difference between grade and expectation and the same expectation. It measures predictive capacity, independently from over or under estimation. The second index, DIFF, is just the absolute difference between the grade and the expectation. It assumes, therefore, a negative value when the expectations are higher than the real evaluations and there is overconfidence. So it measures the existence and magnitude of overconfidence. Table 2 defines all variables used in the following analysis.

Table 2: List of variables

BETTER= 1 if SCORE2> SCORE1; 0 otherwise
DIFF1= SCORE1 - EXP1
DIFF2= SCORE2 - EXP2
EXP1: pre-test expected score of the first test
EXP2: pre-test expected score of the second test
FEM= 1 if student is female; 0 otherwise (females are the 68% of the sample)
OVERC1 = 1 if EXP1 > SCORE1 (overconfidence); 0 otherwise
OVERC2 = 1 if EXP2 > SCORE2 (overconfidence); 0 otherwise
PREC1= (SCORE1-EXP1)^2 /(EXP1)
PREC2= (SCORE2-EXP2)^2/(EXP2)
SCORE1: score of the first test (on 30)
SCORE2: score of the second test (on 30)
SCORE2OVERC= SCORE2, if the student is overconfident in the second exam; 0 elsewhere
SCORE2UNDERC= SCORE2, if the student is not overconfident in the second exam; 0 elsewhere
SCOREHIGH: score got at the end of the high school (from 0.6 to 1)
TEST1AGAIN = 1 if the student choose to sustain the recover test; 0 elsewhere

[3] The two data set are different, also because the academic situations are different. In my class expectations were collected a week before the exam and not two days before it. All of my students were attending their first year at University and were all nineteen or twenty years old. No ethnic variables can be tested here, as all students were Italian. In respect to Grimes's paper no data are available on the average number of hours spent studying. Only students who attended at least three quarters of the lessons were allowed to keep the two mid term examinations and so my sample is homogeneous also in this regard. So I don't need to check for students' age, ethic origins, or presence at lesson. If some variables are missing in respect to Grimes (2002), *new* variables are available and this allow to widen his analysis. Besides, in this case there was an attempt to reduce overconfidence, motivating students with the additional grade in case of good forecasting.

My analysis will be focused on the following points.

- Knowing more

Previous evidence on the effect of this variable is not so clear. While worst performances are generally associated with overconfidence and bad calibration, best results are not necessarily linked with fine forecasting capacities, as sometimes they are associated with underconfidence. If evaluating one's one knowledge is difficult, and expectations tended to converge toward a mean value identical for all students, best students would be underconfident and worst one overconfident. This would depend on a common incapacity to evaluate. A similar empirical relation between overconfidence and performance could arise also in a different way: overconfidence could be the cause of a bad preparation and not (only) its effect. Students who feel optimistic in relation to their goals might have a minor stimulus to study and get, therefore, worse results. These two causes can be both at stake at the same time and cannot be distinguished here.

The capacity to calibrate could be a kind of individual attitude: persons with good meta-cognitive capacities are always better calibrated. The capacity to have realistic expectations could also be related to a specific domain: I'm well calibrated where I'm expert, but not where I'm not. To test for these alternative ideas, we need different indicators of capacity. Students' capacities can, in fact, be measured in different ways: we can use the performance in the same exam on which we test overconfidence or a different and previous index of attainment (as Grimes, 2002 does). If we measure students' performance with the same grade used to compute PREC and the presence of overconfidence, there could be measurement problems related to mathematical effect. Assume, for example, that EXP= SCORE + u, where u is a random error (so expectation would be equal to the real score but for a random error), we would have:

$$prec = \frac{[score-(score+u)]^2}{(score+u)} = \frac{[score-score-u]^2}{(score+u)} = \frac{[u]^2}{(score+u)}$$

In this case PREC would reduce, when SCORE increases - as the expected error would be fixed - reducing the relative difference between EXP and SCORE. In this case the link would have a limited economic meaning: in relative term, best students' errors are less relevant than those of worst ones. Different measures of students' capacities (as score got in a previous exam) have a similar problem as they could be related with the score in the same exam, just acting as its proxy.

Under different hypothesis on the determinants of expectations, the relation between PREC and SCORE can be more informative but, also, again biased by mathematical effect. If all students state a similar expectation at a mean level, best ones would be underconfident and the other overconfident, just for a technical matter.

So analysing the links between calibration and capacities is not so easy. There are different possible explanations and some possible technical difficulties, also for the simple link between overconfidence and performance. In this paper I try to distinguish between the competence in the specific exam and a more general habit to good result. The performance in the specific exam is mainly used as a control variable, so to be able to analyse the role of the other effect. As a secondary effect, yet, my analysis will anyway allow to gather additional empirical evidence on this point. The questions to be answered are these one:

- are the worst students overconfident?
- are even the best students underconfident and not so well calibrated?

To answer them, especially the last one, it is necessary to evaluate if the relation between score and predictive calibration is non linear and/or if there is a linear link between score and DIFF.

- Past experience

Ceteris paribus, students used to get good grades could be more overconfident and also less calibrated in their expectations (overconfidence is necessarily related to a wrong calibration). Usually, as seen, competence should allow a good forecasting performance or, at least, could determine underconfidence. Here I'm stating that, leaving a part capacity and preparation,

students with better curricula can be more overconfident in their expectations that those used to get lower scores.
This idea can be tested using the score got in the high school as an indicator of students past habits to good or bad results. Such a variable is probably correlated with the score got in the two tests under consideration. The correlation is, yet, not necessarily strong, as there are many differences between high school and university. Besides, students arrived from various high schools, and a same score might be the results of quite different levels of real study, work and effective attainment. It could, therefore, be a better indicator of my student's habit to good or bad results than of their real capacities. Its effect on the new performance can be ambiguous and change according to the other regressors inserted in the estimations.

- Improving the score

Grimes's students who improved their grade between their first and second exam, had a better predictive calibration and were less overconfident. I'll try to test this idea, using the dummy variable BETTER.

- Sex

In many empirical analyses, females result less overconfident. I'll test this effect too (with the dummy variable FEM).

- Satisfaction for score1

My students, if not satisfied with their first term grade, had the chance to repeat it a second time. These students need to be signalled in the estimations, as they took a new exam and eventually got a new score, who substituted SCORE1. I use the dummy variable TEST1AGAIN, to control for this aspect. A student, who re-did the first test, probably had a higher self evaluation which, ceteris paribus, could create overconfidence. Among many similar students those who decided to give again the exam, are the ones with a higher self evaluation, so, probably also more self-confidence.

5. Results

5.1. Preliminary statistics

Table 3 shows some preliminary statistics for the main indicators under analysis. In both tests, mean expectations were higher than the real grades. In the first one, the difference is about more than four points and almost all students were overconfident (93% of the sample). In the second test there was a strong improvement, but still 56% of the participants were overconfident and their mean expected grade was still higher than the real one. The score increased and the expectations decreased, with an obvious following reduction in their difference (DIFF). Table 3b shows that all of the differences between the first and second exam are significant according both to the t and the non parametric test for paired data.

Table 3a. Descriptive statistics for the main indicators

	Minimum value	Maximum value	Mean	Standard Deviation
SCOREHIGH	0.60	1	0.87	0.12
EXP1	19	29	26.3	1.75
SCORE1	16.5	27.0	22.1	2.88
PREC1	.00	2.89	.9	.78
DIFF1	-8.50	1.00	-4.3	2.44
EXP2	18	28	23.6	1.99
SCORE2	17	29	22.7	3.49
PREC2	.00	2.89	.47	.57
DIFF2	-8.50	5.00	-.86	3.25

Table 3b. Evolution of the main indicators between the first and second test

			Pvalue t- test	Pvalue
EXP1-EXP2	26.3	23.6	.000	0.000
SCORE1	22.1	22.7	.084	0.053
PREC1	.91	.47	.000	0.000
DIFF1	-4.25	-.86	.000	0.000

Table 4 shows some correlation coefficients useful for the present analysis. All values are positive, as expected, so SCOREHIGH is positively related with both expectations and test performances, but the relation is not so strong (even if it is not so easy defining how high they should be). The expectations for both tests are positively and significantly related with the real score, but again the correlation seems relatively weak.

Table 4. Correlation coefficients among some selected variables

	EXP1	SCORE1	EXP2	SCORE2	SCOREHIGH
EXP1	1				
SCORE1	.54 .000	1			
EXP2	.58 .000	.60 .000	1		
SCORE2	.41 .001	.63 .000	.40 .002	1	
SCHOREHIGH	.51 .000	.48 .000	.43 .001	.54 .000	1

N=57 in all cases; the first number in any cell is the parameter, the second is the P-value of the t test

5.2. Estimations

- Predictive calibration in the first test (PREC1)

Table 5 shows the results of a series of estimations aimed at explaining the determinants of predictive calibration in the first test. As almost all students were overconfident, PREC1 measures just the dimension of the overconfidence (and it makes no sense estimating also DIFF1). When SCOREHIGH is inserted without SCORE1 (and therefore it works as a proxy for student capacity), the sign of its parameter is negative and significant. When we add SCORE1 in the model, SCOREHIGH is still significant but takes a positive sign. In this case SCORE1 assumes a negative sign: higher scores are related with lower values of PREC1. The last estimation (number 5), includes among the regressors, the squared value of SCORE1 so to test for eventual non linear effects of the score. This variable is significant and positive, while SCORE1 remains negative. The overall relation between SCORE1 and PREC1 is therefore represented by a parabola. In a first moment, when the score increases, PREC decreases, but then it starts to increase too. So the best students are not as well calibrated as those with a medium score. Also in this last estimation, SCOREHIGH maintains its significance and a positive value. The dummy control for female students (FEM) is never significant.

Table 5. Estimations with PREC1 as dependent variable

	est1	est2	est3	est4	est5
Costant	2.3 .003	2.3 .003	5.6 .000	5.2 .000	13.0 .001
FEM		.1 .562	.2 .251	.1 .403	.2 .194
SCOREHIGH	-1.6 .068	-1.7 .059		1.1 .076	1.3 .032
SCORE1			-.2	-.2	-1.0

			.000	.000	.005
SCORE1^ 2					.02 .031
R^ 2	0.06	0.07	0.65	0.67	0.70
Adjusted R^ 2	0.04	0.03	0.63	0.65	0.67
P-value F	0.068	0.16	0.00	0.00	0.00

N=57 in all cases; the first number in any cell is the parameter, the second is the P-value of the t test

- Predictive calibration in the second test (PREC2)

The first ten columns in table 5 show a series of estimations with PREC2 as dependent variable. The first one includes as regressors: FEMALE, SCORE1, SCOREHIGH, and TEST1AGAIN. In this model, SCORE1 is supposed to get the effect of student's capacity, while SCOREHIGH should keep its past experience. SCORE1 has a negative effect: its higher values increase the capacity to calibrate, while high SCOREHIGH reduces it. The second estimation is qualitatively similar, but with SCORE2 (a better indicator of the performance in the second test) instead of SCORE1. Results are similar but the overall fit is improved. If both scores are inserted (est3), SCORE1 looses its significance, while SCOREHIGH is always positive and significant. This result holds in all different estimations. So its effect seems quite robust. Estimations 5 and 6 include PREC1, which can be seen as a measure of both capacity (worse students had higher values of PREC1) and of an individual attitude to wrong calibration and overconfidence; it is highly significant and it gives the estimation a strong fit. TEST1AGAIN has always a positive and significant effect. The parameter of the dummy variable for females is almost always significant and negative; yet it loses its significance in the last estimation (est9), when SCORE22 (the squared value of the score got in the second exam) is inserted, and, therefore, the model account for the non linear effect of the score. Probably among medium-scored students there are many females. When the model keeps the difference between medium and high scores, FEM lose most of its effect[4].
The fit of my models (measured by the adjusted R-squared) is about the same as the one estimated by Grimes (equal to 0.31), but when the SCORE22 is introduced in the estimation, it increases quite a lot. This shows the relevance of this non linearity. As in Grimes, even here students who improved their grade in the second test (BETTER=1) had a better capacity to calibrate; yet, the variable which keeps this effect ceases to be significant, when SCORE22 is inserted in the estimation, maybe because students who improved were the ones with medium votes (it's difficult, in fact, to improve an already high grade); so when the model can account for the predictive un-calibration of high graded students, BETTER become no more relevant. Overconfident students (OVERC2=1) are not worst calibrators than others; this is another hint for the existence of problem in calibration arising from underconfidence.

- Absolute difference between score and expectation in the second test (DIFF2)

The last five columns of table 6 show estimations with DIFF2 as dependent variable. When positive, this variable signals underconfidence (SCORE2>EXP2), while it is negative when there is overconfidence. The first two estimations include respectively SCORE1 or SCOREHIGH to measure student's capacities. They both have a positive and significant effect. Again SCORE1 is proposed together with TEST1AGAIN, which is still significant (here and in all other models): students who did again the first exams are more overconfident. When SCORE2 is inserted in the model (est13), both SCORE1 and SCOREHIGH acquire a negative and significant sign, while SCORE2 has a positive effect. If expectations were randomly selected, we would just have a positive sign (if students were not capable of forecasting their grades, the ones with a better performance would be less overconfident by construction). Yet, as seen, EXP2 and SCORE2 are significantly correlated and

[4] Even PREC1 - in an estimation not shown - loses its significance when SCORE22 is inserted, as it probably keep, in some way, the non linear effect of the score.

expectations are, therefore, not completely random. If all students had the same capacity to forecast their grade, this parameter should be equal to zero. So a positive sign is not completely obvious and signals a general difficulty in forecasting. The last estimation tries to find out if SCORE2 plays a different effect for overconfident students, in respect to the others. The effect of SCORE2 is now divided in two variables, one (SCORE2OVERC) for overconfident students, the second for all others. So the effect of the score is now allowed to be different for these two groups. If overconfidence arises just because of bad knowledge, the parameter of SCORE2UNDER should be null or negative (best students should have the lowest values of DIFF2) while the other parameter should be positive. Yet, on the contrary, the two parameters are almost identical (and therefore are not significantly different). Even the best students were bad calibrators, because of underconfidence. When controlling for all variables (est13-15) the dummy for females (FEM) results significant with a positive sign: female students are less overconfident (while they are not necessarily well calibrated in their expectations, as seen).
When inserted with SCORE2 in the estimation of PREC2, SCORE1 is not significant, while it is significant in the estimation of DIFF2. So a good performance in the first exam didn't influence predictive calibration in the second test, but increased the probability of overconfidence.

- Two more estimations

In the estimation of DIFF2 which includes SCORE2OVERC and SCORE2UNDER, SCOREHIGH is almost irrelevant. Probably there is a relation between the score got at high school and being overconfident. If this effect is signalled by the two different variables for SCORE2, SCOREHIGH loses significance. So SCOREHIGH has a strong impact on PREC2 and on the probability of being overconfident, while its effect on DIFF2 is weaker. This suggests the need for estimations 10 and 15. The first one uses PREC2 as dependent variable and includes the squared value of SCOREHIGH, to account for an eventual non linear effect (suggested by these empirical findings, while un-previously expected). The estimation confirms such non linearity. So, after a given threshold, a high value of SCOREHIGH starts to reduce overconfidence.
The model 15 uses two variables instead of SCOREHIGH, allowing two different effects for overconfident (SCOREHIGHOVERS) and for the others (SCOREHIGHUNDER) students. The sign of the parameter is negative and significant for the overconfident persons, while it is non significant for the others, as now expected.

Table 6. Estimation with PREC2 or DIFF2 as dependent variable

	PREC2										DIFF2				
	est1	est2	est3	est4	est5	est6	est7	est8	est9	est10	est11	est12	est13	est14	est15
Costant	.8 .190	1.0 .068	1.1 .072	1.1 .067	-.5 .358	-.5 .389	1.3 .072	8.5 .030	11.8 .000	7.3 .030	-7.4 .020	-8.4 .007	-11.7 .000	-9.2 .000	-8.7 .000
FEM	-.3 .032	-.3 .016	-.3 .017	-.3 .020	-.3 .016	-.3 .017	-.3 .030	-.3 .073	-.2 .131	-.2 .165	.9 .276	.6 .542	.9 .035	.8 .036	.7 0.051
SCORE1	-.06 .043		-.0 .732	-.0 .566			-.1 .006	-.8 .033		-0.0 0.983	.3 .046		-.3 .000	-.3 .003	-.2 0.007
SCORE1^ 2								.02 .054							
SCOREHIGH	1.2 .085	1.7 .008	1.8 .008	1.8 .008	1.2 .048	1.2 .049	1.8 .012	1.8 .007	1.5 .007	15.4 .038		8.3 .021	-5.2 .016	-3.1 .117	
TEST1AGAIN	1.0 .001	.8 .002	.8 .002	.8 .004	.8 .006	.8 .006	.8 .005	.8 .003	.7 .005	.7 .004	-4.3 .007		-2.2 .011	-2.0 .012	-2.2 0.003
SCORE2		-.1 .000	-.1 .003	-.1 .098					-1.1 .000	-1.2 .000			1.0 .000		
SCORE2^ 2									.02 .000	.02 .000					
BETTER				-.1 .639	-.3 .034	-.3 .057	-.4 .024	-.3 .032	-.0 .970	-.1 .721					
PREC1					.3 .002	.3 .003									
OVERC2						-.0 .814	-.1 .617								
SCORE2(OVERC)														.7 .000	.8 0.000
SCORE2(UNDER)														.7 .000	.5 0.000
SCOREDIP^ 2										-8.3 .059					
SCOREHIGHOVER															-6.5 0.006
SCOREHIGHUNDER															2.9 0.335
R^ 2	0.28	0.39	0.39	0.40	0.39	0.39	0.37	0.34	0.58	0.61	0.23	0.11	0.81	0.85	0.87
Adjusted R^ 2	0.23	0.35	0.34	0.33	0.33	0.31	0.29	0.41	0.53	0.55	0.18	0.08	0.79	0.83	0.85
P-value F	0.002	0.000	0.000	0.000	0.00	0.001	0.001	0.000	0.00	0.00	0.04	0.04	0.000	0.000	0.00

N=57 in all cases; the first number in any cell is the parameter, the second is the P-value of the t test

- Logit estimations

Table 7 shows the results of three logit estimations with OVERC2 as dependent dichotomous variable (it is equal to one for the overconfident students, while it is equal to zero for all others, that is both underconfident students and those who perfectly forecasted their grades). When inserted with SCORE2, both SCORE1 and SCOREHIGH have a positive and significant sign, increasing the chance of overconfidence. SCORE2 has a negative sign: better students are less probably overconfident. SCORE1 alone is significant, but with a positive sign, as in this case (est2) it mainly measures students' capacity. SCOREHIGH alone is negative but not significant. The dummy variable for female students is never significant.

Table 7. Logit estimations with OVERC2 as dependent variable

	est1	est2	est3
Costant	3.2 .123	4.3 .068	12.3 .035
FEM	-.1 .813	-.3 .586	-1.4 .220
SCORE_HIGH	-3.3 .170		13.4 0.037
SCORE1		-.2 .085	.5 .062
TEST1AGAIN		20.9 .999	20.2 .999
SCORE2			-1.4 .000
Chi-squared	0.330	0.039	0.000
correct prediction	63%	60%	83%

6. Comments and conclusions

My analysis tried to bring more empirical evidence on overconfidence. I evaluated if my students were able to state a realistic expectations for their future performance. Most of my students were overconfident in predicting their future performances in both of their exams. As they were at their beginning at the University, their difficulties in forecasting could be not surprising, while it is less obvious their overconfidence. In the second test there was, yet, a significant improvement in their predictive capacity and self evaluation, but overconfidence still prevailed. In both cases, overconfidence in expectations was connected with students' performance, while predictive calibration showed a non linear relation with their grade. Also better students, in fact, weren't well calibrated. This evidenced a general incapacity to predict one's own performance. Students who got the highest grades were, ceteris paribus, less calibrated than the ones who got medium scores. So while bad performances were characterized by overconfidence, the best performances come along with underconfidence. If expectations were fixed and identical for all students, this result would just be obvious: nobody is capable of making realistic forecasting and given identical expectations, better students are underconfidence while worse ones are overconfidence (yet while should worse students have the same expectations as the best ones if not because of overconfidence?). Expectations had a lower variance than the real grades but they were not identical for all students and resulted positively correlated with the real grades. Better students had higher expectations that worse one and proportionally higher grades. A path of this kind could be, at least in part, the cause and not the effect of student's performance as it could reveal a different attitude to meta-cognition and/or self confidence which would stimulate a more attentive study.

When controlling for the performance in the specific exam, past good results (both the score got in high school - and the score of the first test, when we analyze the second one) stimulated biased expectations, increasing the chance to be overconfident and not well calibrated. Usually, in experiments on overconfidence, there is no measure of the past performance in similar environment, while the score got in high school is here a good indicator of student's habit. So, this is probably the most innovative results of this work, that was also suggested by the literature on rule based behaviour and perception. Persons used to get good results, even if capable of understanding that their knowledge is not so good, could be biased by a kind of inertia in their expectations: "all my past test were always good, maybe even when I wasn't so optimist". According to Gigerenzer (1991), when asked to evaluate their overall performance in a test, persons would probably use their past performances in situations perceived as similar. Previous performances at high school could be a natural reference for students in University. Yet, as in Italy there are many kinds of high school and University requires different skills, this reference could be biased and determine overconfidence. Students used to get good results in bad schools, were probably not able to evaluate that the situations changed and their capacities had been overestimated in the past years. So the relation between capacities and overconfidence was here reverted, and still non linear. The threshold where the effect is reversed is, yet, quite high (at about 0.96); practically it affects only the values of the students who got the maximum grade at high school. This effect could also suggests that best students who got the maximum grade at high school could have some specific meta-cognitive capacities increasing their calibration. Yet, ceteris paribus, students who got the maximum score (100%) was less calibrated than the ones who had a low result (as even 80%). So, better students, ceteris paribus, are more overconfident, yet, usually, better students are also more prepared in the specific exam, so this effect can be hardly measurable and the inertia in expectations could be masked by competence. My analysis, in fact, also evidenced measurement difficulties in studying overconfidence. I proposed different specifications of my models, just to investigate the robustness of the effects under analysis. Some parameters are very robust, while others seemed to be quite dependent on specification. Some variables are, probably, proxies for other ones, and this sometimes hide their real effects, which emerge just when the right variables are also inserted. So, even in the study of overconfidence we need to be careful, so to avoid an excessive trust in our results!

References

- Bandura, A 1982. Self-efficacy: mechanism in human agency. American Psychologist 37:122-147.
- Bandura, A 1989. Human agency in social cognitive theory. American Psychologist 44:1175-1184.
- Bateson G. (1972), *Steps to an Ecology of Mind. Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*, References are based on Bateson (2000), University of Chicago press, Chicago
- Bengtsson C., Persson, M. Willenhag P. (2005) Gender and overconfidence, Economics Letters, vol 86-2, February, 199-203
- Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases." In W. Stroebe and Hewstone M. (Eds.), European Review of Social Psychology (Vol. 2, pp. 83-115). Chichester, England: Wiley.
- Griffin, D & Tversky, A 1992. The weighing of evidence and the determinants of confidence. Cognitive Psychology 24:411-435.
- Grimes, P. W. (2002) The overconfident principles of economics student: An examination of a. metacognitive skill, Journal of Economic Education 33 (1): 15-30
- Hayek (1969), The Primacy of the Abstract. In New Studies in Philosophy, Politics, Economics, and the History of Ideas. Chicago: University of Chicago Press.
- Heiner, R. A. (1983) The Origin of Predictable Behavior, *American Economic Review*, 73 (4): 560-595
- Klayman J., Soll J.B., Gonzalez-Vallejo C, Barlas S. (1999), Overconfidence: It depends on How, What and Whom you Ask, Organizational Behavior and Human Decision Processes, Vol. 79(3), September, pp. 216-247
- Levinthal, D.A. and March, J. G. (1993). The Myopia of Learning.Strategic Management Journal, Vol. 14, 95-112.
- Lichtenstein S. and Fischhoff, B 1977. Do those who know more also know more about how much they know? Organizational Behaviour and Human Performance 20:159-183.
- Livingston J.A. (1997), Metacognition: An Overview, mimeo, http://www.gse.buffalo.edu/fas/shuell/CEP564/Metacog.htm (last download on february the 8th, 2009)

- Ochse C. (2001), are positive self-perceptions and optimistic expectations really beneficial in an academic context?, progressio, vol 23-2, 52-60
- Rabin, M. and J. Schrag (1999) First Impressions Matter: A Model of Confrmation Bias., Quarterly Journal of Economics, February 1999, 114 (1), pp. 37-82
- Rizzello S. (1999), *The Economics of the Mind*, Edward Elgar, Aldershot
- Russo, J.E., Schoemaker, P.J.H. (1992), "Managing overconfidence", Sloan Management Review, Winter, pp.7-17
- Svenson, O. (1980). Are we all less risky and more skillful than our fellow drivers? Acta Psychologica, 47, 143-148
- Van Den Steen E. (2004), Rational Overoptimism (and Other Biases), American Economic Review, 94(4), 1141-1151
- Vanberg V. (1993), "Rational Choice, Rule-Following, and Institutions -- an Evolutionary Perspective," in Mäki U., Gustafsson B., Knudsen K. (eds.), *Rationality, Institutions, and Economic Methodology*, Routledge, London
- Vromen, J. (2004), *Routines, genes and program-based behaviour*, Papers in economics and Evolution, Max Planck Institute, pp. 1-30