Economics Bulletin

Volume 29, Issue 3

Incentives to learn calibration: a gender-dependent impact

Marie-pierre Dargnies
Paris School of Economics, Université Paris 1

Guillaume Hollard
Paris School of Economics, CNRS

Abstract

Miscalibration can be defined as the fact that people think that their knowledge is more precise than it actually is. In a typical miscalibration experiment, subjects are asked to provide subjective confidence intervals. A very robust finding is that subjects provide too narrow intervals at the 90% level. As a result a lot less than 90% of correct answers fall inside the 90% intervals provided. As miscalibration is linked with bad results on an experimental financial market (Biais et al., 2005) and entrepreneurial success is positively correlated with good calibration (Regner et al., 2006), it appears interesting to look for a way to cure or at least reduce miscalibration. Previous attempts to remove the miscalibration bias relied on extremely long and tedious procedures. Here, we design an experimental setting that provides several different incentives, in particular strong monetary incentives i.e. that make miscalibration costly. Our main result is that a thirty-minute training session has an effect on men"s calibration but no effect on women"s.

We are very grateful to Michèle Cohen, Jean-Christophe Vergnaud, Maxim Frolov, Gilles Bailly, Natacha Raffin, Victor Hiller and Thomas Baudin. We are grateful to numerous seminar participants at the JEE conference in Lyon, especially Glenn Harrison, and at the University of Paris 1 and Brown University.

Citation: Marie-pierre Dargnies and Guillaume Hollard, (2009) "Incentives to learn calibration: a gender-dependent impact", *Economics Bulletin*, Vol. 29 no.3 pp. 1820-1828.

Submitted: Apr 16 2009. **Published:** July 28, 2009.

1 Introduction

In the past decades Economists and Psychologists documented a long list of biases , i.e. substantial and systematic deviations from the predictions of standard economic theory [1]. Many economists will argue that these biases only matter if they survive in an economic environment. In other words, if correct incentives are provided subjects should realize that they are making costly mistakes and then change the way they make such decisions in further decision tasks. In this paper we test this claim regarding a particular bias, namely miscalibration. We then create an experimental setting that provides a lot of incentives (decisions have monetary consequences, successful others can be imitated, feedbacks are provided, repeated trials are used, etc). Finally, we test in a subsequent decision task whether subjects still display some miscalibration.

What is miscalibration and why is it important to economists?

Calibration is related to the capacity of an individual to choose a given level of risk. In a typical experiment designed to measure miscalibration, subjects are asked to provide subjective confidence intervals. For example, if the question is "What was the unemployment rate in France for the first trimester of 2007?" and the subject provides the 90% confidence interval [7%,15%], it means that the subject thinks that there is a 90% chance that this interval contains the correct answer. A perfectly calibrated subject's intervals should contain the correct answer 90% of the time. In fact, a robust finding is that almost *all* subjects are miscalibrated. On average, 90% subjective confidence intervals only contain the correct answer between, say, 30% and 50% of the time [2]. Glaser et al. (2005) found an even stronger miscalibration using professional traders.

Miscalibration is a bias having important economic consequences, since miscalibrated people suffer losses on experimental markets (Bonnefon et al., 2005; Biais et al., 2005). Furthermore, it is likely that such a pathology affects the behavior of real traders acting on real markets. Therefore, it does make sense for economists to try to reduce miscalibration and to study the best incentives to do so.

Several psychologists have used various techniques to reduce miscalibration (Pickhardt and Wallace, 1974; Adams and Adams, 1958; Lichtenstein and Fischhoff, 1980), with little success so far.

This paper proposes to provide a maximum of incentives to reduce miscalibration. The main result is that our experimental setting succeeds in reducing overconfident miscalibration but only for males.

The remainder of the paper is organized as follows. Section 2 presents the experimental design. Section 3 presents the results while section 4 discusses them and provides some concluding remarks.

[1] A list of almost a hundred of such biases can be found at
http : *//en.wikipedia.org/wiki/List_of_cognitive_biases*

[2] see Lichtenstein and Fischhhoff (1977) for a survey and (Klayman et al., 1999) for variables that affect miscalibration

2 Experimental design

The experimental subjects were divided into two groups. The subjects of the first group attended a training session and then performed a baseline treatment aiming at measuring their miscalibration according to the standard protocol. The principle of this training session is to offer a whole set of experimental incentives that enhance learning (monetary incentives, tournament, feedback, loss framing). The second group, the control group, performs the baseline treatment only. Since there is no simple incentive scheme that rewards correct calibration in the standard calibration task [3], we chose to consider a task similar to the calibration task in which we can provide the necessary incentives. This task, described in the following section aims at making the subjects realize they have a hard time calibrating the level of risk they wish to take. After having completed this training task, subjects have to complete a standard calibration task for which we only provide incentives for the following evaluation of how subjects did in the calibration task as in Cesarini et al. (2006). A control group who did not go through the training task also completed the calibration task to enable us to measure the effect of the training task.

2.1 The training period

In the training period, the participants were asked to answer a set of twenty questions: ten questions on general knowledge followed by ten questions on economic knowledge.

The set of questions used in the training period was composed of ten questions some of which were used in Biais et al. (2005)'s experiment plus 10 questions on economic culture. In this training period, the subjects were provided with a reference interval for each question that they could be 100% sure the correct answer belonged to. Subjects had to give an interval included in the reference interval. Each player received an initial endowment of 2000 ECUs (knowing that they would be converted into euros at the end of the experiment at the rate of 1 euro for 100 ECUs) before beginning to answer the questions but after having received instructions. They were told that 100 ECUs were at stake for each one of the twenty questions resulting in a loss framing. The payoffs are expressed in experimental currency (ECU). The payoff rule applied for each question was the following :

$$\text{payment} = \begin{cases} -100 * \frac{\text{width of the interval provided}}{\text{width of the interval of reference}} & \text{if the correct answer belongs to the interval provided} \\ -100 & \text{otherwise} \end{cases}$$

According to this formula, the payoff is maximal and equal to 0 when the interval provided by the subject is a unique value, this value being the right answer to the question. In this

[3] Think, for example, of an incentive scheme that would pay a high reward if the difference between the required percentage of hit rates, say 90%, and the actual hit rate (measured over a set of 10 questions) is small. A rational subject can use very wide intervals for 9 questions and a very small one for the remaining question. He is thus certain to appear correctly calibrated, while he is not.

case, the subject keeps the total 100 ECUs at stake for the question considered. If the subject provides the reference interval and consequently takes no risk at all, he loses the 100 points at stake for the question. There is therefore a trade-off between the risk taking and the amount of ECUs a subject could keep if the correct answer fell inside his interval. High risk taking is rewarded by a small loss in the case where the answer belongs to the interval provided. Conversely, a subject who only takes little risk will only keep a few ECUs even if the correct answer does belong to his interval.
Subjects received feedback providing them with the intervals chosen by all the participants (including themselves) ranked by width from the narrower to the wider as well as the payoff corresponding to each interval. They could infer from this feedback whether they had taken too much risk compared to the others. They could also see the ranking of everybody's score after each question so as to trigger a sense of competition.

After they had answered all 20 questions, subjects received general feedback about the first step of the experiment.
People being miscalibrated, we expected them to realize it when they saw that the correct answer fell outside their interval less or more often than they had expected, which resulted in a loss of money. As a result, we expected them to better adjust the level of risk they wished to take for the next questions.

2.2 The standard calibration task

In the next stage, the subjects who had participated in the training period were asked to answer a set of ten questions (five questions on general knowledge followed by five questions on economic knowledge) by giving their best estimation of the answer and then by providing 10%, 50% and 90% confidence intervals. Subjects in the control group had to complete the same task. Before the beginning, subjects were explained in detail what were 10%, 50% and 90% confidence intervals. They were also told that they would receive remuneration regarding this task but that they would only know how the remuneration was established later. As in Cesarini et al. (2006), since it is impossible to find an incentive-compatible payoff scheme for providing confidence intervals [4], their remuneration for the calibration tasks depended on the evaluation the subjects were asked to make afterwards of their and the average subject's performance during the calibration task. There was no feedback between the questions.

3 Results

The experiment took place at the laboratory of experimental economics of the University of the Sorbonne (Paris 1) in July 2007. 87 subjects, most of whom were students, participated in the experiment. 53 students went through the training period before they completed the calibration task, while the control group was composed of 34 subjects. The average earning was 11.16 euros. On average, subjects earned 10.62 euros including a 5 euros show-up fee in the control group and 14.24 euros (8.42 for the training period and 5.82 for the calibration

[4] see footnote 3

task) with no show-up fee for the trained group. One can notice that the payoffs for the calibration task are very similar for the control and the trained group (respectively 5.62 and 5.82 euros). Nevertheless, remember that these earnings do not correspond to how well calibrated participants are but to their ability to predict ex post how well they were calibrated. In consequence, the fact that earnings are very similar across treatments does not mean that subjects did not learn to calibrate better.

3.1 General results on calibration

We find that the subjects from the control group exhibit a high level of miscalibration. Indeed, a lot more than one correct answer out of ten belong to the 10% intervals while fewer than five correct answers out of ten fall inside the 50% confidence intervals and far fewer than nine correct answers out of ten fall inside the 90% intervals. The average hit rates in the control group at the 10%, 50% and 90% levels are respectively 2.03, 3.32 and 4.81. T-tests show that the observed hit rates significantly ($p<0.001$ for the 3 tests) differ from the expected hit rates (respectively 1, 5 and 9 at the 10%, 50% and 90% levels).

At the 10% level, people are found to be under-confident, meaning that they provide too wide intervals. As a result, the correct answer belongs too often to the 10% intervals. This result was expected by Cesarini et al. (2006). At the 50% and 90% levels conversely, subjects display overconfidence as their intervals are too narrow, this is all the more the case for 90% confidence intervals (in line with the results of Glaser et al. (2005)).

Comparing the level of miscalibration we get to those found in other studies,

A surprising feature is that, when asked to evaluate how many correct answers belong to their intervals, the average answers are respectively at the 10%, 50% and 90% levels: 3.47, 5.56 and 8.04 for the control group. Subjects exhibit overconfidence for the calibration task, thinking that they were more cautious than they actually were. Let us, nevertheless, observe that subjects do predict that their calibration is far from being perfect, otherwise their evaluations would have been 1, 5 and 9.

These results indicate that not only are people unable to adjust the width of their intervals to the risk level indicated (they are miscalibrated) but they are also unable to predict their bias correctly (they are over or underconfident).
To sum up, people seem to overestimate their underconfidence and underestimate their overconfidence.

3.2 The effect of training on miscalibration and confidence in calibration

Trained subjects have only slightly higher hit rates at the 10%, 50% and 90% level than subjects from the control group. The differences in hit rates between the control and the trained group are not significantly different at any reasonable level. [5]

[5]The hit rates are respectively at the 10%, 50% and 90% levels 2.03, 3.32 and 4.81 for the control group and 2.40, 3.80 and 5.33 for the trained group. To get an idea of levels of miscalibration found in other

We find that the median 10% interval width is larger for the trained group than for the control group for 7 questions out of ten. For the 3 remaining questions, the median width of intervals is equal across treatments. Note that this goes in the sense of a worsening of the underconfident miscalibration observed at 10% as people tend to provide too wide intervals at 10%. One reason why we may find such a result is that subjects may not consider the underconfident miscalibration as a bias and consequently, they may not try to correct it.

The same result is found when we compare median widths of 50% intervals (wider intervals in the trained group than in the control group for 7 questions, the reverse for 1 question and equal median intervals across treatments for the 2 remaining questions). As for 90% intervals, for six questions out of ten the interval width is larger for the trained group while the control group provided wider intervals than the trained group for 1 question. [6]

It may be interesting to study the link between the "theoretical" distribution of hit rates of a perfectly calibrated subject (who has a 90% chance for an answer to fall into any of his 90% confidence intervals...) and the one we actually observe. We report two figures showing the theoretical and actual distributions of 90% intervals hit rates for women and men. Those figures make miscalibration very prominent. We then ran a two-sample median test, separately for women and men, on the distributions of hit rates in the control and the training groups. We find that our training has a significant effect on men's 90% calibration (p=0.089) while no significant effect is found for women. Men's 90% calibration is improved by our training which can be seen on figure 1 by the shift in the distributions of hit rates between the control and the training treatments. No effect is found for miscalibration at the 10% and 50% levels.

We ran logistic regressions of the dummies ICA10 ("the 10% interval contains the correct answer"), ICA50 and ICA90 on the same variables (see Table 1). We observe that the treatment significantly increases the probability for the correct answer to fall in the 50% and 90% intervals provided for almost all of the questions (the interaction terms between the questions and the treatment are always positive and almost always significant). It is true but to a smaller extent for the 10% intervals. If anything, our treatment seems to make subjects provide wider intervals (even if this result is far from always reaching significance)

studies, notice that Russo and Schoemaker (1992) obtained hit rates at the 90% level between 4.2 and 6.2, while Klayman et al. (1999) found 4.3. However, the level of miscalibration is obviously very sensitive to the set of questions used. Since half the questions we used were taken from Biais et al [2005], we can compare the level of miscalibration we found to those of that study. Using no incentive, the average 90% hit rate in their study is 3.6 while we find respectively 4.8 and 5.3 in our control (where subjetcs no they will get a payment but have to wait until the end of the calibration task to find out how it will be calculated) and training (where subjects are in the same situation and previously went through the training period) group. It therefore seems like the presence of incentives does increase hit rates.

[6] If we compare average interval widths, which seems less relevant as averages are sensitive to extreme values, we find that for 7 (6) questions out of ten the average width of 10% and 50% (respectively 90%) intervals are larger for the trained subjects, while for the remaining 3 (respectively 4) questions, the opposite is true.

As variances of interval widths are often very different across the control and trained group and as a way of eliminating the influence of extreme values, we ran a Wilcoxon-Mann-Whitney test. We found that the 90% interval widths are significantly different (either at the 1%, 5% or 10% levels) for 5 questions out of ten while 10% and 50% intervals widths are significantly different respectively for 3 and 6 questions out of ten.

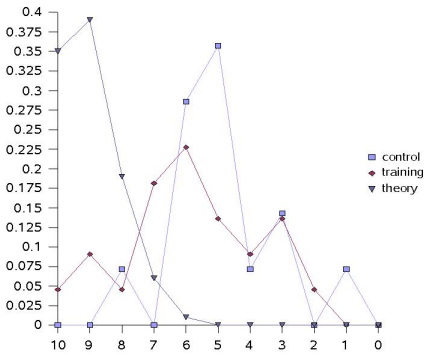

Figure 1. Theoretical hit rates and actual hit rates of men from control and trained group.

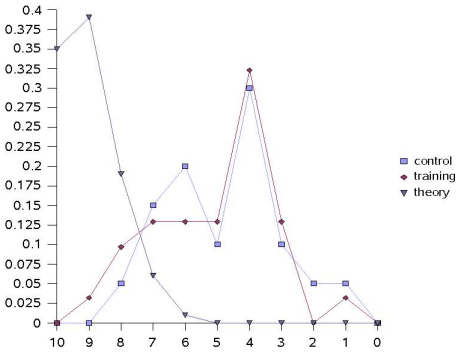

Figure 2. Theoretical hit rates and actual hit rates of women from control and trained group.

and it significantly helps subjects catch the correct answer in their confidence intervals more often. Consequently, the incentives we provided during a short training period decrease the overconfident miscalibration we observe for the 50% and 90% intervals but, to a smaller extent, makes the underconfident miscalibration noticed for the 10% intervals worse.

Table I. Logistic Regression of "the 10, 50, 90% interval contains the correct answer" (ICA10 ICA50 ICA90)

Variable	ICA10	ICA50	ICA90	Variable	ICA10	ICA50	ICA90
Intercept	-0.7402	0.4416	1.3924	q8	-1.1865	-1.8656	-1.4668
	(0.2799)	(0.4948)	(0.0392)		(0.0656)	(0.0019)	(0.0171)
Sexe	0.0412	0.3924	0.0738	q9	-1.7349	-2.3641	-2.3985
	(0.8915)	(0.1391)	(0.7711)		(0.0190)	(0.0002)	(0.0002)
treatment	-0.5951	-1.5940	-1.1804	q10	-1.7381	-1.9947	-1.9420
	(0.2896)	(0.0048)	(0.0485)		(0.0188)	(0.0011)	(0.0020)
sextreatment	0.1398	-0.0941	0.2821	q2t	0.8536	1.7735	0.0282
	(0.7124)	(0.7764)	(0.3801)		(0.2290)	(0.0152)	(0.9776)
Age	0.00302	0.00324	-0.00832	q3t	0.8063	2.5884	1.5503
	(0.9108)	(0.8895)	(0.7117)		(0.3400)	(0.0018)	(0.0441)
Education	0.0507	0.0773	0.0350	q4t	0.7633	2.1252	1.5607
	(0.4971)	(0.2327)	(0.5801)		(0.2941)	(0.0032)	(0.0345)
q2	0.2337	-0.3636	1.2972	q5t	0.7561	2.3399	1.6269
	(0.6766)	(0.5381)	(0.1441)		(0.4115)	(0.0014)	(0.0279)
q3	-1.3719	-2.8147	-2.3541	q6t	-0.7081	1.2329	0.6230
	(0.0447)	(<.0001)	(0.0003)		(0.4120)	(0.1041)	(0.3992)
q4	-0.4377	-1.5713	-1.3416	q7t	1.3228	2.1902	1.2944
	(0.4481)	(0.0074)	(0.0291)		(0.0935)	(0.0034)	(0.0833)
q5	-1.7691	-1.8656	-1.4668	q8t	1.0284	2.2127	1.5161
	(0.0167)	(0.0019)	(0.0171)		(0.1977)	(0.0026)	(0.0404)
q6	-0.7728	-1.8656	-1.3416	q9t	1.0319	2.3218	2.5139
	(0.1985)	(0.0019)	(0.0291)		(0.2509)	(0.0027)	(0.0010)
q7	-1.1124	-1.9534	-1.7502	q10t	1.5328	2.3593	1.6794
	(0.0856)	(0.0015)	(0.0052)		(0.0804)	(0.0015)	(0.0246)

Note: p-values are in brackets.

This general picture masks some strong heterogeneity across subjects. We can control for several sources of heterogeneity. However, the gender variable captures almost all of it. We observe indeed that there is virtually no improvement in women's calibration especially when we compare the median hit rates between the treatments while men increase their median hit rate by 0.5 point at the 50% level and by 1 point at the 10% and 90% levels.

The difference in interval width between the control and the training treatment seems to be larger for men than for women, indicating that men learned more than women to reduce

their overconfidence. Using a Wilcoxon-Mann-Whitney test, we find that 10% confidence intervals are significantly wider for the trained group respectively for five questions out of ten and zero question out of ten for men and women. Let us notice that in the trained group both men and women had more than one correct answer inside their 10% intervals exhibiting underconfident miscalibration. As a result, an increase of 10% intervals causes an aggravation of underconfidence. For 50% intervals, the width increases significantly between the control and the training treatments respectively for two and six questions out of ten. Finally, concerning 90% intervals, the difference is significant in three cases and four cases out of ten respectively for women and men.

4 Discussion and conclusion

This paper contributes to a literature interested in cognitive biases having economic consequences. We focus on miscalibration, a very robust bias correlated with losses on experimental financial markets and bad entrepreneurship.

In line with the existing literature on miscalibration, our subjects strongly suffer from the miscalibration bias, their 50% and 90% intervals being too narrow (overconfident miscalibration). We find that subject's 10% intervals are too wide (underconfident miscalibration). These results are widespread in the population according to the literature and there are very few exceptions. Furthermore, subjects overestimate their underconfidence and underestimate their overconfidence. The fact that people overestimate their underconfident miscalibration could mean that they do not consider it as a bias. Maybe being too cautious is seen as a good thing. Previous attempts to reduce miscalibration relied on very long and repetitive training periods.

Our thirty-minute training punishing miscalibrated behavior by money losses results in an improvement of calibration at the 50% and 90% levels but the underconfident miscalibration observed at the 10% level is made worse by the training. Some consequences can be drawn. It is unlikely that miscalibration disappears in a market environment, since we provided the kind of incentives that are expected on real markets. According to our results, real traders are likely to underestimate the risk they take when they think they invested in a very secure asset. Symmetrically they take less risks than they think when they invest in risky assets. So, the overall effect of miscalibration on real markets is ambiguous. Furthermore, we find that men's calibration can be improved by our training period, while women's cannot. The incentives we implemented had no effect on women. There are probably many incentives one could think of that would have a differential effect on men and women. Women traders may need either a longer training period which would give them more time to get rid of their miscalibration or a different kind of incentives they would react to.

References

Adams, P. and J. Adams (1958). The effects of feedback on judgmental interval predictions. *American Journal of Psychology 71*, 747–751.

Biais, B., D. Hilton, K. Mazurier, and S. Pouget (2005). Judgmental overconfidence, self-monitoring and trading performance in an experimental financial market. *Review of Economic Studies 72*, 287–312.

Bonnefon, J., D. Hilton, and D. Molian (2005). A portrait of the unsuccessful entrepreneur as a miscalibrated thinker. Working Paper.

Cesarini, D., O. Sandewall, and M. Johannesson (2006). Confidence interval estimation tasks and the economics of overconfidence. *Journal of Economic Behavior and Organization 61*, 453–470.

Glaser, M., T. Langer, and M. Weber (2005). Overconfidence of professionals and laymen: Individual differences within and between tasks?

Klayman, J., J. Soll, C. Gonzalez-Vallejo, and S. Barlas (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes 79*(3), 216–247.

Lichtenstein, S. and P. Fischhhoff (1977). Do those who know more also know more about what they know? *Organizational Behavior And Human Performance 20*, 159–183.

Lichtenstein, S. and B. Fischhoff (1980). Training for calibration. *Organizational Behavior And Human Performance 26*, 149–171.

Pickhardt, R. and J. Wallace (1974). A study of the performace of subjective probability assessors. *Decision Sciences 5*, 347–363.

Russo, J. and P. Schoemaker (1992). Managing overconfidence. *Sloan Management Review 33*, 7–17.