How To Undo Biased Self-Assessments

Marian Krajč, Andreas Ortmann, Dmitry Ryvkin

30 September 2008
This is an unedited draft

Abstract

We report the results of two experiments (one field, one laboratory) through which we examine the impact of general information and specific information on the quality of self-assessment ("calibration") in various tasks and feedback conditions. We find a strong positive effect of naturally available information (both general and specific) on calibration in the field experiment. We also identify a positive effect of specific information separately in the laboratory experiment. Maybe unsurprisingly, in both experiments it is the unskilled who improve their calibration most. Our results suggest that the unskilled may not be doomed to be unaware (if indeed they are).

JEL Classification: C46, C91, C93, D01, D81, D83, D84
Keywords: calibration, judgement errors, unskilled, unaware, metacognition, experiment

1. Introduction

The so-called *unskilled-and-unaware problem* was first identified by psychologists Kruger and Dunning (1999). The authors conducted, with students, several experiments in which subjects were asked to estimate their relative standing (ranking) and absolute performance (score) in various tasks. The authors identified, both for reported rankings and scores estimates, three regularities: (i) people ranked at the bottom of the skills distribution overestimated their ranking/score; (ii) those ranked at the top of the skills distribution underestimated their ranking/score; (iii) these miscalibrations were typically highly asymmetric in that many more unskilled overestimated their ranking/score and often did so quite dramatically. Kruger and Dunning (1999) argued that the unskilled also lack the metacognitive ability to realize their incompetence and were thus afflicted by a double-curse. The term *unskilled-and-unaware problem* typically refers to the large miscalibration (*overconfidence*) of the unskilled, and implicitly to the asymmetry in miscalibration between the unskilled and the skilled.

The claim drew significant attention. As we write this article, scholar.google documents more than 400 citations for Kruger and Dunning (1999). Not surprisingly, a number of follow-up studies were subsequently conducted and various explanations provided (e.g., Krueger and Mueller 2002; Burson, Larrick, and Klayman 2006; Krajc and Ortmann 2008). In a recent article, Ehrlinger, Johnson, Banner, Kruger, and Dunning (2008) addressed the concerns of some of their critics and asserted that the unskilled-and-unaware problem remains alive and well.

The problem of miscalibration, or biased self-assessments, is also of eminent interest to economists. Camerer and Lovallo (1999), for example, showed that people are largely overconfident when entering laboratory markets with payoffs depending on their relative skills. Their paper, too, has garnered significant attention. As we write this article, scholar.google documents more than 300 citations for Camerer and Lovallo (1999).

In the current manuscript, we report the results of an experimental investigation of the impact of various types of information on miscalibration in situations – such as in Kruger and Dunning (1999) and Camerer and Lovallo (1999) – where subjects, for

reasons that are in dispute (e.g., Krajc and Ortmann 2008), may start out with biased self-assessments. Our study – a field experiment in which we embedded a laboratory experiment – contributes to a better understanding of the circumstances under which biased self-assessments might occur, and how they can be undone. In passing, we address a question that Juslin, Winman, and Olsson (2000) identified as being in need of an answer: what, if any, is the relationship between calibration in general-knowledge tasks and calibration in skill-oriented tasks, the two major paradigms of overconfidence studies in psychology.

Subjects in our two experiments demonstrate, on average, initially overconfident behavior. We show, however, that information improves calibration, especially of the unskilled. This reduces the unskilled-and-unaware problem over time, as conjectured in Krajc and Ortmann (2008). We also identify an interesting design problem that seems inherent in overconfidence studies and that neither our study nor other studies so far have successfully addressed. Our results on the relationship between calibration in general-knowledge tasks and calibration in skill-oriented tasks do not warrant strong conclusions and remains a question in need of an answer.

The present paper is organized as follows. First, we briefly review the literature concerned with the unskilled-and-unaware problem and related issues such as the issue of representative stimuli. In the third section, we motivate and detail our research objectives and research strategy. In the fourth section, we discuss the design and implementation of the experiments. In the fifth section, we present the results. In the last section we discuss our results and the design problem mentioned earlier.

2. The unskilled-and-unaware problem and related issues – a brief literature review

As mentioned, the results and conclusions of Kruger and Dunning (1999) did not go unchallenged. For example, Krueger and Mueller (2002) contested those results when they showed that the use of the unreliable measures[1] causes the measured ability to regress toward the mean, which induces overestimation (underestimation) in the lower

[1] "Unreliability" of a measure denotes the imperfect correlation between a predictor variable and a criterion variable. In this case, estimated percentiles are not a perfectly reliable measure of real percentiles. Lack of reliability in a test makes the highest (poorest) performers look less (more) able than they are.

(upper) part of the distribution.[2] In addition, to explain the asymmetry, the authors used the presence of the better-than-average effect (the belief by a majority of people that they are better than the average). This strikes us as a less than persuasive argument since it makes the explanandum into the explanans.

Burson, Larrick, and Klayman (2006) introduced task difficulty into the unskilled-and-unaware problem. Specifically, they demonstrated that the degree of over- and underconfidence depends on the task difficulty. Indeed, their results were similar to those in Kruger and Dunning (1999) for easier tasks (with asymmetry in over- and underestimation). For harder tasks, Burson and her colleagues found, in contrast, less overestimation of the unskilled and more underestimation of the skilled. In fact, asymmetry in over- and underconfidence disappeared (or even was reversed – more underestimation among the skilled than underestimation among the unskilled) in experiments with harder tasks.

Ehrlinger, Johnson, Banner, Kruger, and Dunning (2008) addressed some of the questions about the earlier work of Kruger and Dunning (1999) by using financial and social incentives and real-world situations; they also controlled for the unreliability of measures. Despite these changes in experimental implementation, they replicated the pattern observed in Kruger and Dunning (1999) of overestimation of their skills by the unskilled and underestimation of their skills by the skilled, with miscalibration much more dramatic for the unskilled than the skilled in all treatments. Ehrlinger et al. (2008) also searched for the cause of the identified miscalibration in percentile ranking. Using counterfactual regression analysis, they explored how more accurate information on own scores and the scores of others would have improved calibration in rankings. It was found that, overall, knowing their own score would have helped subjects more than knowing others' scores, but the two types of information would have had approximately the same effect on the skilled. Thus, it was shown that the skilled and the unskilled are differentially affected by information.

Krajc and Ortmann (2008) offered an alternative explanation of the unskilled-and-unaware problem: They constructed a simple model that shows that the unskilled, rather

[2] Several authors have proposed models of miscalibration based on subjects making random errors in judgment. For example, Erev, Wallsten, and Budescu (1994) created a general model that assumed that subjects' confidence was a function of true judgments and error.

than being more unaware than the skilled, face a tougher inference problem which, at least partially, explains the alleged lack of metacognitive ability of the unskilled. Their model is based on two assumptions. First, they claim that students' skills[3] – at least in the studies of Kruger and Dunning (1999) and Ehrlinger et al. (2008) – have a bounded J-distribution.[4] Second, the authors assume that the self-assessment process involves unsystematic noise.[5] Thus Krajc and Ortmann (2008) use computational simulations to generate patterns of miscalibration similar to those reported by Kruger, Dunning, and their collaborators. The results suggest that the unskilled may indeed be unaware but that it is not necessarily for a (relative) lack of metacognitive skills.

Krajc and Ortmann (2008) also discuss the conditions under which they expect the unskilled-and-unaware problem to disappear. They conjecture, as do Camerer and Lovallo (1999), that feedback is likely to ameliorate the miscalibration problem.

Camerer and Lovallo (1999) introduced skill-dependent rankings and rank-dependent payoffs into a market entry experiment and showed that people enter excessively into laboratory markets, suggesting overconfidence. This result has been contested by Elston, Harrison, and Rutstroem (2006) who showed that neither non-entrepreneurs nor entrepreneurs were overconfident about their skills in market entry games. In contrast, the wannabe-entrepreneurs in their sample were.

The contradictory results of Camerer and Lovallo (1999) and Elston et al. (2006) suggest that Camerer and Lovallo (1999) may not have had enough relevant control variables such as measures of risk aversion and desire to win which Elston et al. (2006) used; they also suggest that the specifics of the subject pool matter. Of particular interest is the finding that entrepreneurs seem to have been calibrated reasonably well. This finding is in line with results on the performance of experts in domains as diverse as

[3] Students are typically used in studies of the unskilled-and-unaware problem.

[4] A J-distribution is a distribution with a monotonically decreasing convex pdf describing a relatively large number of the unskilled and a decaying upper tail for the skilled. Krajc and Ortmann (2008) argue that the samples used in earlier studies quite likely satisfied this assumption; they also show that this assumption can to some extent be weakened and that it is easy to account in their framework for legacy cases (i.e., applicants who happen to be children of alumni and therefore might be admitted even when their grades might not warrant it), etc.

[5] This assumption is often used in the literature (e.g., Erev, Wallsten, and Budescu, 1994). Krajc and Ortmann (2008) argue that noise is likely to be correlated with familiarity (and hence feedback about one's own standing) with a particular domain. "If one does not know one's relative standing in a particular context, one is likely to use one's self-assessment from other domains as a proxy, which adds to the error." (Krajc and Ortmann, 2008)

weather forecasting (e.g., Murphy and Winkler, 1984), horse race betting (e.g., Johnson and Bruce, 2001), and games of skill and chance such as bridge (e.g., Keren, 1997).

In our view, the issue of calibration is also fundamentally linked to the issue of representativeness of stimuli in experiments (e.g., Gigerenzer, Hoffrage, and Kleinbolting, 1991; Dhami, Hertwig, and Hoffrage, 2004). Juslin, Winman, and Olsson (2000) showed, with a meta-study based on 130 data sets, that the so-called hard-easy effect (overconfidence more common for samples with hard questions and underconfidence more common for samples with easy questions) in general-knowledge tasks appears typically only in studies that use non-representative stimuli (e.g., selected alternatives for comparison), as does overconfidence. Usage of non-representative stimuli offers an interesting explanation of the results – the dependence of calibration on task difficulty – in Burson, Larrick, and Klayman (2006). However, this explanation is not uncontroversial, as Klayman, Soll, Gonzáles-Vallejo, and Barlas (1999) showed that the degree of overconfidence varied across domains, yet was not a function of domain difficulty.

Importantly, Juslin et al. (2000) identified the relationship between calibration in general-knowledge tasks and calibration in skill-oriented tasks – the two major paradigms of overconfidence studies in psychology – as one in need of an (empirical) answer.

3. Motivation and research hypotheses

It is our main goal to study the impact of various types of information on miscalibration (the unskilled-and-unaware problem). Krajc and Ortmann (2008) have conjectured that (lack of) information about own ability and abilities of others plays an important role in miscalibration; indeed there is some evidence from related judgment and decision tasks that supports this conjecture (e.g., Duffy and Hopkins, 2005; Engelmann and Strobel, 2000).

We distinguish two kinds of information, *general information* and *specific information*. General information is here understood as all the natural information that is generally accessible to all participants. In our experiments, this includes gossip, inferences from study group impressions, or in-class contributions of others. This

information can be thought of as noisy, less reliable or "soft." In contrast, specific information is here understood as the information that is given to participants directly and privately. In our experiments, it is the information about the students' past performance (both absolute and relative), which can be thought of as "hard" information, or explicit "feedback."

Our research hypotheses regard the impact of the two types of information on miscalibration:

Hypothesis 1: *General (soft) information decreases miscalibration.*

Hypothesis 2: *Specific (hard) information decreases miscalibration even more*.

To test these hypotheses, we conducted two experiments (Experiments 1 and 2). Experiment 1 was a field experiment: we compared midterm and final exam predictions of a newly constituted class in a real-world setting. This experiment addressed primarily the *general information hypothesis*. Although specific information was also present in Experiment 1, it was impossible, due to the nature of the field situation under study, to disentangle its impact from that of general information. In turn, Experiment 2, a laboratory experiment that was embedded in the field experiment, addressed both hypotheses.

Note that both experiments were not marred by subject selection problems as our participants were "pseudo-volunteers" (see Eckel and Grossman, 2000).[6] A possible disadvantage with pseudo-volunteers is that the subjects may simply not be interested in participating in the experiment (Harrison and Rutstroem, 2007, especially fn. 79). In light of the time our experiments took and the substantial financial incentives we provided, as well as our observation of our pseudo-volunteers' conduct, we do not believe that we have to worry much about this possible disadvantage.

4. Experimental design and implementation

4.1. Experiment 1 (the field experiment)

[6] We hasten to admit that, although our subjects were pseudo-volunteers, they were highly selected along some dimensions, and the selection process that brought them to the experiment (wanting to become a Ph.D. student at CERGE-EI) was related to experimental tasks (writing micro exams and adding numbers). Studying the effects of that selection was, of course, part of what we were interested in (e.g., Krajc and Ortmann 2008). More on this below.

Each year, CERGE-EI in Prague (Czech Republic) invites selected students from Central European countries, and countries further East, to the preparatory semester (prep), and then admits the best of them for graduate studies based on their results in the prep. Prep students are likely to have been among the best in their college classes in their home countries. When students arrive at CERGE-EI, they have minimal information about the abilities of others (although they might anticipate what kind of people has been invited to the prep semester and although they might understand that the quality of the education systems from which the class is recruited differs widely). Prep students represent a suitable subject pool for investigating the issue of self-assessment under incomplete information (as regards composition of the sample) as well as increasingly more complete information (acquired over time).[7]

During the prep semester, students typically take four courses: microeconomics, macroeconomics, mathematics, and English (academic writing). In each of the four courses, they have a midterm exam, a final exam, and regular homework. For our field experiment, we have used students' self-assessments and performance in the microeconomics course.

Specifically, in Experiment 1 we asked students of the microeconomics course to predict their performance[8] both on an absolute (score) and relative scale (percentile), in the midterm and final exams. Students made these predictions twice for the midterm exam (in week 1 of the preparatory semester, and in week 5 right before the midterm exam), and once for the final exam (in week 9 right before the exam).[9] Appendix C illustrates the timeline of Experiment 1.

A total of 49 (respectively, 52) students made their predictions about midterm performance in week 1 (in week 5), and 45 (51) of these students participated in the midterm exam.[10] A total of 46 students made their predictions about the final

[7] We did a pilot experiment with prep students in 2004. We identified the unskilled-and-unaware problem in the data. The magnitude of the effect decreased toward the end of the semester in the pilot study, too.
[8] Ferraro (2005) used in-class exams to study the relationship between incompetence and overconfidence. He concluded that overconfidence in inversely proportional to competence (performance). Thus, he essentially confirms the results of Krueger and Dunning (1999).
[9] Complete instructions to Experiment 1 are available at http://home.cerge-ei.cz/krajc
[10] Altogether 53 students wrote the midterm exam but 2 people came to the exam after the questionnaires with predictions had been collected.

performance in week 9, and 45 of them wrote the final exam.[11] For each question, the participant with the best prediction was paid 500 CZK.[12] All participants were told that their predictions would not affect their grades and that no one but the researchers would see the data.

Since Experiment 1 was realized at three different points in time (weeks 1, 5, and 9), it allows us to observe the evolution of the level of miscalibration over time under the influence of general information alone (weeks 1-5) and general and specific information jointly (weeks 5-9). At each point in time we measured calibration in each subject's own score ("What is your prediction of your own score on the midterm [final] exam in microeconomics?" – henceforth referred to as "score") and percentile rank ("What do you think is the percentage of people in the group who will perform better than you on the midterm [final] exam in microeconomics?" – henceforth referred to as "percentile"). Over time, students could be assumed to acquire general information about their absolute and relative standing in the microeconomics course and other courses. Additionally, they received specific information about their relative and absolute standings after the midterm exam.

The first question allowed us to measure over- and underestimation of subjects' own ability. With the second question we measured, like Kruger and Dunning (1999), percentile ranking. Experiment 1 featured a real-world setting with high stakes (at least for prospective Ph.D. students coming from Central Europe and further East) and natural information (both of the general and specific kind). We did not give our subjects any artificial feedback; they only received the natural information communicated during such a course: homework grades, midterm results, distribution of midterm scores, etc. They also received indirect feedback from other classes and communication with their peers.

4.2. Experiment 2 (the lab experiment embedded in the field experiment)

[11] Altogether 46 students wrote the final exam. One student came to the exam after the questionnaires with predictions had been collected. One student did not submit the final exam booklet. There was one student who wrote the final but did not write the midterm exam.

[12] At the time 20.50 CZK was equal to $1 and the average hourly wage was approximately 100CZK. Thus, payments were clearly non-trivial.

Experiment 2 was a laboratory experiment conducted in two stages. In each stage, we used two tasks (for the timeline and general structure see Appendix C).[13]

Task 1: Participants had to sum, within a 3-minute time limit, sets of five 2-digit numbers without the use of calculators (see also Niederle and Vesterlund, 2007, and Bruegger and Strobel, 2007). This task is a skill-oriented task (mathematical skill).

Task 2: Participants had to answer, within a 2-minute time limit, a quiz containing 20 two-alternative general knowledge questions (widely investigated in psychology).[14] In Stage 1, we asked for a comparison of the population of pairs of European Union countries ("Which of the following two countries has a larger population?") while in Stage 2, in order to avoid learning, we asked for a comparison of the population of pairs of the 50 most populated countries in the world.[15] This task is a general-knowledge task (knowledge of geography).

A total of 49 (respectively, 45) students participated in Stage 1 (Stage 2) of this experiment. Stage 1 (Stage 2) lasted 25 (20) minutes. All participants were paid according to their performance in the experiment. The average payoff was 177 CZK (313 CZK) in Stage 1 (Stage 2).

Stage 1 was conducted during week 1 of the preparatory semester and Stage 2 at the end (week 9) when students could be assumed to have more information about their relative standing in the group (hypothesis 1: "*general information*"). We did not tell our subjects during Stage 1 that Stage 2 would follow. All instructions were read aloud.

In Stage 1, after providing a brief general introduction to the experiment, we asked our subjects to fill in a short questionnaire (age, gender, and background – mathematician or economist). Then we continued with the instructions. We explained the first task – summing 5 two-digit numbers – and gave an example. The subjects were informed that for each correctly solved problem they would be paid 5 CZK. Afterwards, we distributed sheets with 22 summing problems and gave our subjects 3 minutes to solve as many of these problems as possible. Then, we asked subjects to provide estimates of their own score ("How many summing problems do you think you solved

[13] Complete instructions are available at http://home.cerge-ei.cz/krajc.
[14] E.g., for review see Juslin, Winman, and Olsson (2000).
[15] By learning we mean that some people, motivated by Stage 1 of the experiment, could learn the population of the EU countries and thus we would artificially change the knowledge and might get non-representative data.

correctly?"), and of their percentile ranking ("What do you think is the percentage of people in the group who performed better than you?"). Subjects providing the most accurate estimates to each of these questions were paid 500 CZK. Subjects were informed about this earnings possibility beforehand.[16]

We then explained the second task: comparing population sizes of pairs of European Union countries. The subjects were rewarded with 5 CZK for each correct comparison of 20 pairs of countries they did within 2 minutes. Subsequently, we asked them again to provide estimates of their absolute and relative performance. These questions allowed us, in analogy to the questions in experiment 1, to measure over- and underestimation of subjects' ability and percentile ranking.

Stage 2 was similar to Stage 1, with the following changes. First, we increased the incentives to 10 CZK for each correctly solved summing problem in task 1.[17] Second, we changed the reference class in task 2 and used the 50 most populated world countries instead of the countries of the European Union. Third, in task 2 we gave our subjects 40 general knowledge questions keeping the reward for correct answer constant (5 CZK), thus effectively doubling the incentives similarly to task 1. Fourth, in Stage 2, one half of the participants received for each task feedback about their absolute and relative performance at stage 1 (own score, percentile, and the group average score). Subjects for the feedback treatment had been randomly selected (in a stratified manner)[18] just before each task. Therefore, in addition to some indirect (natural) feedback acquired from the micro, macro, and math results from the midterm exams and homework, some subjects received specific feedback. The specific feedback allowed us to investigate how the strength of the feedback influences calibration (Hypothesis 2).[19]

[16] We understand that students could intentionally under-perform to improve their predictions. To minimize this effect, we told students that ties in predictions will be broken according to performance. We did not see, or find in the data, any evidence of intentional underperformance.

[17] Because the time gets scarcer towards the end of the semester we decided to increase the incentives for our subjects. We doubled the reward for correct answers in Stage 2 of Experiment 2. The analysis of the high and very high payoff treatments in Rydval and Ortmann (2004) suggest that this increase does not matter in any significant way.

[18] In line with their performance in Stage 1, we sorted subjects into four quartiles and randomly selected half of the subjects in each quartile for the feedback treatment.

[19] In addition, we asked our subject to provide, together with the score and percentile estimates, also predictions/estimates of the average score in each stage of each experiment. Finally, we asked participants of Stage 2 of Experiment 2 to provide additional predictions for minimum and maximum scores achieved in each task.

In Experiment 2, unlike in Experiment 1 where stimuli materials were given by the instructor of the class, we used tasks that allowed us better control for the representativeness of stimuli. First, we clearly specified the classes of questions (the so-called reference classes: all two-digit numbers, all countries in the European Union, the 50 most populated countries in the world). Second, we randomly chose the numbers and country pairs from the corresponding reference classes. Thus, we presented our subjects with representative samples of problems as suggested by previous research (e.g., Gigerenzer, Hoffrage, and Kleinbolting, 1991; Dhami, Hertwig, and Hoffrage, 2004; Juslin, Winman, and Olsson, 2000).

Incentives play an important role in various types of studies (see Camerer and Hogarth, 1999; Rydval and Ortmann, 2004). In Experiment 2, we used tasks that are responsive to higher effort (e.g., for general knowledge employing more cues, as suggested by Gigerenzer, Hoffrage, and Kleinbolting, 1991) and therefore we expect that monetary incentives will increase the accuracy of given answers and thus also the measured ability. To motivate the subjects to give answers as precise as possible, we used a linear incentives scheme.

Because of the number of participants, we had to make a choice between using 2 feedback treatments or 2 incentives treatments. The evidence in Cesarini, Sandewall, and Johannesson (2006) suggests strongly that, at least in the present context, incentives are of lesser importance than feedback. We therefore decided to use two feedback conditions, which is not ideal but was the best we could do under the circumstances.

Note that task 1 in Experiment 2 is more skills oriented while task 2 is more knowledge oriented. Because of the different nature of the tasks, we are able to observe how the distributions of skills and knowledge differ and how they are related to each other, if at all.

6. Results

In this section we report, and discuss, summary statistics and hypotheses tests for Experiments 1 and 2. Tables and figures referred to in this section can be found in

Appendix A and Appendix B, respectively. The timeline and structure of the experiments are summarized in Appendix C.

6.1. Experiment 1

To recall, Experiment 1 involved subjects making three predictions about their absolute and relative performance in two exams. Midterm prediction 1 (M1) was collected in week 1 of the preparatory semester; midterm prediction 2 (M2) was collected in week 5 right before the midterm exam; final prediction (F) was collected in week 9 right before the final exam.

Table 1 shows the summary statistics for the subjects' actual performance ("Actual Score") and predictions ("Predictions"). All scores have been rescaled to the 0-100 range for comparability. Subjects made predictions for their own score ("Score") and the percentage of subjects with a score higher than theirs ("Percentile"). As seen from Table 1, subjects, on average, exhibit overconfidence: the average predicted percentiles are significantly smaller than 0.5. Additionally, subjects are miscalibrated in terms of absolute performance (scores) although one could argue that that measure is confounded by the fact that it is ultimately the instructor who determines the task difficulty and grading, thus in this context this measure is of arguably less value.

We describe miscalibration by two measures: (i) overestimation, defined as the difference between the predicted and the real value of the corresponding score,[20] and vice versa for percentiles;[21] and (ii) absolute deviation, defined as the absolute value of the difference between the predicted and the real value.

Overestimation in scores describes subjects' overconfidence regarding their absolute performance, while overestimation in percentiles describes overconfidence in relative standings. The absolute deviation measure describes miscalibration more generally: even in the absence of significant overestimation, absolute deviation may

[20] For example, if one's own score is 14 and one's estimate of one's own score is 16, then we observe a positive number (2) which means overestimation of one's own score.

[21] In the case of percentiles, a positive number means overestimation of own relative ranking (underestimation of number of percentage of better performing people). E.g., if one's real percentile ranking is 0.2 and one predicted that 0.1 of people will perform better – that person would have a positive number here.

capture considerable miscalibration since over- and underconfidence might cancel each other out.

Table 2 summarizes overestimation and absolute deviation for predictions M1, M2, and F. We formed overestimation and absolute deviation variables for each subject and tested the hypotheses of mean overestimation and mean absolute deviation (MAD) being equal to zero. The resulting means and p-values for the overestimation and absolute deviation variables are given in Table 2. Positive and significant mean overestimation and absolute deviation are observed in all predictions, implying, on average, overconfidence and general miscalibration in absolute and relative performance.

However, as seen from Table 2, overestimation in scores decreases over time from 33.41 (M1) to 29.22 (M2) to 14.28 (F). Similarly, overestimation in percentiles decreases from 0.23 (M1) to 0.20 (M2) to 0.11 (F). The MAD in scores and percentiles also decreases over time, with a less significant change from M1 to M2, and a more significant one from M2 to F. Thus, both overconfidence and general miscalibration are affected by information as predicted. The change from M1 to M2 is due to the general information students obtained from course work and communication with their peers, while the (more substantial) change from M2 to F is due to both general and specific information obtained after the midterm scores and relative standings have been revealed.

Table 3 presents the results of statistical pairwise comparisons of miscalibration in scores and percentiles between M1 and M2, M1 and F, and M2 and F. For each pair of predictions, we compare both the average overestimation and absolute deviation measures of miscalibration using the paired t-test and Cohen's d effect size statistics. The cells where the comparison yields a significant difference are shaded gray.

Between predictions M1 and M2, overestimation in scores decreased at a relatively weak significance level, and the effect size is small ($p=0.139$, $d=0.17$). At the same time, absolute deviation decreased significantly, with a small to medium effect size ($p=0.027$, $d=0.28$). As for miscalibration in percentiles, the decrease in overestimation is small ($p=0.247$, $d=0.096$), while the decrease in absolute deviation is significant, with small effect size ($p=0.030$, $d=0.15$). Thus, although the general information acquired between M1 and M2 leads to a relatively small and insignificant decrease in average

overestimation, it does improve overall calibration significantly both for absolute performance and relative standings.

Between predictions M1 and F, there is a large and highly significant decrease in overestimation of scores (p=0.001, d=0.88), whereas for percentiles the significance level is moderate, and the effect size is medium (p=0.139, d=0.46). The mean absolute deviation in scores decreases significantly, with a large size effect (p=0.000, d=1.06); in percentiles, the decrease in percentiles is also significant, with a medium to large effect size (p=0.042, d=0.61). Predictions M1 and F are separated by the largest time span, so the largest impact of information (both general and specific) on miscalibration is expected in this pair of predictions. Calibration in absolute performance improved significantly, with both average overestimation and the MAD strongly decreasing. Calibration in relative standings improved overall (the MAD decreases strongly), but only moderately in terms of overestimation.

Between predictions M2 and F, overestimation in scores decreases significantly, with a large effect size (p=0.001, d=0.75), while overestimation in percentiles decreases insignificantly, with a medium effect size (p=0.384, d=0.35). The MAD behaves similarly, with a large and significant decrease in scores (p=0.001, d=0.76), and a medium and insignificant decrease in percentiles (p=0.286, d=0.42). Between predictions M2 and F, subjects acquired both general and specific information (midterm scores) about their absolute and relative performance, therefore, as expected, calibration improves stronger than between M1 and M2 in all dimensions.

The prevalence of overconfidence, and the role of information in decreasing miscalibration, is illustrated in more detail by Figure 1. Figure 1 shows the predicted exam scores and percentiles as functions of real scores and percentiles for each of the three predictions. Solid squares (midterm prediction 1), empty squares (midterm prediction 2), and crosses (final prediction) are the actual observations, while the lines (respectively, solid, dashed, and dotted) are obtained from linear regressions of predicted scores on real scores and predicted percentiles on real percentiles (with all regressions including intercept). The regression results are shown in Table 4.

Figure 1 and Table 4 suggest the following observations:

1. Most of the data points are above the 45 degree line for scores and below the 45 degree line for percentiles, indicating overestimation of own scores and overconfidence regarding own relative standing.
2. The slopes of all the estimated lines are smaller than one, indicating that the unskilled overestimate their scores more than the skilled, and are more overconfident regarding their relative standing than the skilled.
3. All estimated lines intersect the 45 degree line, indicating that the most skilled are, in fact, underestimating their scores, and are underconfident regarding their relative standing.
4. The intercept of the estimated dependence of predicted scores and percentiles on real scores and percentiles decreases with time, and the slope increases with time, indicating the role of information.

Overall, as illustrated in Figure 1, the estimated dependence of predicted scores and percentiles on real scores and percentiles becomes closer to the 45 degree line (the ideal calibration) with students obtaining more information. Recall that in the case of exam predictions, students received direct feedback after the midterm exam. However, we observe improved calibration already before this information was revealed – this is most likely based on indirect feedback obtained from course work and interactions with classmates.

Yet another representation of the impact of information on calibration is shown in Figure 2. Here, students are sorted into quartiles according to their performance in the midterm and final exam. For each quartile, Figure 2 shows the evolution of the average predicted percentile from M1 to M2 to F. This presentation mode, which can also be found in Krueger and Dunning (1999) and much of the literature, adds an interesting twist. First, note that while in the top two performance quartiles (1 and 2) there is practically no reduction in miscalibration, the reduction is very strong for the bottom two quartiles. Thus, the unskilled are affected by information significantly stronger than the skilled. Second, although in quartiles 1-3 students become relatively well calibrated by week 9, in the lowest performance quartile 4 there appears to be still large residual miscalibration. Indeed, as seen from Figure 1, no subjects placed themselves in the bottom 20% of the class despite all the general information and feedback they received.

We conjecture that two phenomena could contribute to this result. First, although subjects might well understand that they are at the bottom of their class, they may be hesitant to share that insight with the experimenter because they might fear that such an act of self-assessment – notwithstanding our promise that our data would not be shared with their instructor – might be revealed and affect their grade. Second, and drawing on the arguments proposed by Koeszegi (2006), subjects might just not be willing to accept the fact that they are at the bottom of the class. Our setup and data do not allow us to tease apart these two reasons which identify an interesting design problem of overconfidence studies that neither our study nor other studies so far have successfully addressed.

6.2. Experiment 2

To recall, Experiment 2 was conducted with the same subjects in two stages. Stage 1 (S1) was conducted together with prediction M1 of Experiment 1 (week 1), and Stage 2 (S2) together with prediction F of Experiment 1 (week 9). At each stage, subjects performed two tasks (Task 1 and Task 2), and made predictions regarding their absolute and relative performance after each task.

One key difference between Experiments 1 and 2 is in that in Experiment 1 all predictions are made before the corresponding activity takes place, while in Experiment 2 predictions are made after the fact. We chose this design because the accuracy of predictions has been incentivized with significant amounts of money, and we did not want subjects to intentionally underperform to match their predictions.[22] This problem was of no significance in Experiment 1, where we believe subjects had sufficiently strong incentives to perform well in the exams. Thus, we expect much better calibration in absolute performance in Experiment 2. At the same time, calibration in relative standings should be independent of this difference in designs.

The other key feature of Experiment 2, compared to Experiment 1, is our ability to control specific feedback. During prediction S2, we gave half of the subjects specific feedback about their absolute and relative performance in the corresponding task at Stage

[22] To discourage such behavior even further, we told subjects that ties in most accurate predictions would be broken according to subjects' performance in the task.

1. The subjects receiving feedback were determined randomly, separately for each task, and in a stratified manner. We sorted subjects into quartiles according to their performance at Stage 1, and approximately equalized the number of subjects getting feedback in each quartile. We expect the subjects who received specific feedback to exhibit better calibration at Stage 2 compared to those who did not.

As before, we start with the summary statistics for performance and predictions. Table 5 summarizes the actual scores obtained in each task at Stage 1 and Stage 2 of Experiment 2, as well as the predicted scores and percentiles for each task in both stages.

As seen from Table 5, subjects are overconfident, with average percentile predictions well below 0.5. At the same time, subjects are relatively well calibrated in terms of absolute performance, as expected due to the design feature discussed above. Interestingly, although in Task 1 subjects, on average, exhibit overestimation of absolute performance, they exhibit underestimation of absolute performance in Task 2. This may be due to the fact that Task 2 is a general knowledge task, while Task 1 is a specific skill task, but the difference does not appear to be statistically significant.

Table 6 summarizes overestimation and absolute deviation for predictions S1 and S2 by task. Similarly to Experiment 1, we tested the hypotheses of mean overestimation and the MAD being equal to zero. For predictions S2, we summarize miscalibration overall and separately for the subjects who did and did not receive specific feedback. Interestingly, for Task 1 mean overestimation in scores increases over time from 0.62 (S1) to 1.45 (S2). Similarly, the MAD in scores increases from 1.34 (S1) to 1.73 (S2). However, the subjects receiving feedback experience a smaller increase in miscalibration than those without feedback: mean overestimation of 1.00 (S2 with feedback) versus 1.84 (S2 without feedback) and the MAD of 1.29 (S2 with feedback) versus 2.13 (S2 without feedback). The effect of feedback for Task 1 is even more drastic for percentiles. Miscalibration in percentiles decreases overall, with mean overestimation dropping from 0.11 (S1) to 0.094 (S2), and the MAD dropping from 0.26 (S1) to 0.21 (S2), but the subjects who received feedback exhibit practically no overconfidence, on average, with mean overestimation of 0.018 and insignificant, and much smaller general miscalibration, with the MAD of 0.11, while the subjects without feedback are more miscalibrated than at Stage 1, with mean overestimation of 0.16 and the MAD of 0.30.

For Task 2, there is significant underestimation of 2.28 in scores at Stage 1, which practically disappears for subjects with feedback at Stage 2 (mean overestimation of -0.079 and insignificant), while turns into overestimation (mean overestimation of 0.73, albeit insignificant) for subjects without feedback. The MAD for scores slightly decreases overall, from 2.81 (S1) to 2.63 (S2), and again the effect is stronger with feedback, with the MAD of 2.43 (S2 with feedback) and goes in the opposite direction without feedback (2.85, S2 without feedback). For overestimation in percentiles, the results are similar to those for Task 1. There is a decrease in overconfidence overall, with mean overestimation in percentiles decreasing from 0.15 (S1) to 0.11 (S2), but the decrease is primarily due to the subjects who received feedback. With feedback, mean overestimation at Stage 2 is 0.0018 and insignificant, but without feedback it is 0.20, i.e. larger than in Stage 1. The MAD for percentiles increased overall from 0.26 (S1) to 0.30 (S2), but the increase is less dramatic for subjects with feedback (0.28 at Stage 2) than for those without feedback (0.32 at Stage 2).

Table 7 shows paired t-test and Cohen's d effect size statistics for the comparison of miscalibration in scores and percentiles between Stage 1 and Stage 2 for each task overall, with, and without feedback. Similarly to Table 3, the cells where the comparison yields a significant difference are shaded gray.

In Experiment 2 for both Task 1 and Task 2 we observe a much better calibration in scores than in Experiment 1. This is primarily due to the fact that predictions in Experiment 2 were made after the tasks have been completed, and also due to the more transparent and familiar nature of the tasks. For Task 1, we observe stronger miscalibration in scores over time, even for the subjects who received specific feedback. We believe this is due to learning, which made the inference problem harder. Task 1 is a skill-oriented task, and at Stage 2 subjects used the techniques they learnt at Stage 1 and performed better (mean actual score increased from 6.85 to 7.70, see Table 5). At the same time, they based their predictions at Stage 2 on the results of Stage 1 and could not evaluate their improvement adequately. For percentiles, there is a significant improvement of calibration overall and especially for the subjects who received feedback. For Task 2, calibration in scores improved, changing from strong underestimation at Stage 1 to weak overestimation at Stage 2.

Figure 3, similarly to Figure 1, illustrates how general information acquired between Stage 1 and Stage 2 and specific feedback affected miscalibration in Experiment 2. The predicted scores and percentiles are shown as functions of real scores and percentiles for each task. Solid squares (S1), empty squares (S2 without feedback), and crosses (S2 with feedback) are the actual observations, while the lines (respectively, solid, dashed, and dotted) are obtained from linear regressions of predicted scores on real scores and predicted percentiles on real percentiles (with all regressions including intercept). The regression results are shown in Table 8.

Figure 3 and Table 8 suggest the following observations:

1. Subjects are reasonably well calibrated in terms of scores already at Stage 1 in both tasks. There is no apparent over- or underestimation in scores. This is primarily a consequence of the predictions being made after the task has been completed.
2. Most observations for percentiles are below the 45 degree line indicating overconfidence in relative standings in both tasks.
3. At Stage 2, subjects are more miscalibrated in terms of scores compared to Stage 1. This is primarily due to inadequate assessment of learning in the skill-oriented task, and a change in the reference class of problems in the general knowledge task.
4. For both scores and percentiles, and for both tasks, the subjects who received feedback are better calibrated at Stage 2 than those who did not.

Overall, general information acquired between Stage 1 and Stage 2 does not improve subjects' calibration in Experiment 2, but the subjects who received specific information are calibrated better than those who did not.

7. Discussion and conclusion.

We have reported the results of two experiments (one field, one laboratory) through which we examined the impact of general information and specific information (feedback) on the quality of self-assessment ("calibration") in various tasks and feedback conditions. We find a strong positive effect of general information on calibration in the field experiment (Experiment 1). Recall that in the case of exam predictions, students received direct feedback after the midterm exam. However, we observe improved calibration already before this information was revealed – this is most likely based on

indirect feedback obtained from course work and interactions with classmates. In the lab experiment (Experiment 2) where we could control specific information, the subjects who received specific information were calibrated better than those who did not. These results bear on the debate about the reality of cognitive illusions. At least for the allegedly well-established overconfidence phenomenon – one of the bones of contention in this debate, – it seems to take surprisingly little to reduce miscalibration quite dramatically.

In Experiment 2, subjects were comparatively well calibrated in absolute performance at all stages. This is mainly due to the design feature that, unlike in Experiment 1, subjects made predictions after performance and thus were familiar with the stimuli and could better assess their ability in performing the tasks. Our use of representative stimuli may have also contributed to this result. At the same time, calibration in percentiles in Experiment 2 is not better than in Experiment 1, indicating that, as expected, the timing of predictions does not matter for overconfidence.

Maybe unsurprisingly, in both experiments it is the unskilled who improve their calibration most. Thus, our results suggest that the unskilled may not be doomed to be unaware (if indeed they are).

In none of our experiments we observe anyone with predicted/estimated percentile rank in the worst 20% of the group (Figures 1 and 3 in Appendix B have no data points with predicted percentile higher than 0.8; see also Figure 3). This is an interesting design problem potentially pertaining to all overconfidence studies that neither our study nor other studies so far have successfully addressed.

Our results on the relationship between calibration in general-knowledge tasks and calibration in skill-oriented tasks do not warrant strong conclusions.

References

Brueggen A., Strobel M., (2007). Real effort versus chosen effort in experiments, *Economics Letters, 96,* 232–236.

Burson A.K., Larrick P.R., & Klayman J. (2006). Skilled or Unskilled, but Still Unaware of It: How Perceptions of Difficulty Drive Miscalibration in Relative Comparisons. *Journal of Personality and Social Psychology, 90,* 60-77.

Camerer D., & Hogarth R.M., (1999). The effects of financial incentives in experiments: A review and capital-labor-production theory, *Journal of Risk and Uncertainty, 19:1-3,* 7-42.

Camerer C., & Lovallo D., (1999). Overconfidence and Excess Entry: An Experimental Approach, *American Economic Review, 89:1,* 306-318.

Cesarini D., Sandewall O., & Johannesson M., (2006). Confidence Interval Estimation Tasks and the Economics of Overconfidence, *Journal of Economic Behavior & Organization, 61 (3),* 453-470.

Dhami K.M., Hertwig R., & Hoffrage U., (2004). The Role of Representative Design in an Ecological Approach to Cognition, *Psychological Bulletin, Vol. 130, No. 6,* 959–988.

Duffy, J., Hopkins, E., (2005). Learning, information, and sorting in market entry games: theory and evidence. Games and Economic Behavior, 51, 31–62.

Eckel C.C., & Grossman P.J., (2000). Volunteers and pseudo-volunteers: The effect of recruitment method in dictator experiments. *Experimental Economics, Vol. 3, Num. 2,* 107-120.

Ehrlinger J., Johnson K., Banner M., Kruger J., & Dunning D. (2008). Why the Unskilled are Unaware: Further Exploration of (Absent) Self-Insight Among the Incompetent. *Organizational Behavior and Human Decision Processes*, 105 (1), 98-121.

Engelmann, D., Strobel, M., (2000). The False Consensus Effect Disappears if Representative Information and Monetary Incentives Are Given. *Experimental Economics, 3,* 241–260.

Elston J.A., Harrison G.W., & Rutstroem E.E., (2006). Characterizing the Entrepreneur Using Field Experiments, incomplete draft.

Erev I., Wallsten T.S., & Budescu D.V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101,* 519-527.

Ferraro J.P., (2005). Know Thyself Incompetence And Overconfidence. *Experimental Laboratory Working Paper Series* #2003-001, Dept. of Economics, Andrew Young School of Policy Studies, Georgia State University. Revised January 2005.

Gigerenzer G., Hoffrage U., & Kleinboelting H., (1991). Probabilistic Mental Models: A Brunswikian Theory of Confidence. *Psychological Review, 98,* 506-528.

Harrison G.W., & Rutstroem E.E, (2007). Risk Aversion in the Laboratory. *Working Paper 07-03,* Department of Economics, College of Business Administration, University of Central Florida.

Johnson J., & Bruce A., (2001). Calibration of Subjective Probability Judgments in a Naturalistic Setting. *Organizational Behavior and Human Decision Processes, 85,* 265-290.

Juslin P., Winman A., & Olsson H. (2000). Naïve Empiricism and Dogmatism in Confidence Research: A Critical Examination of the Hard-Easy Effect. *Psychological Review, 107,* 384-396.

Keren G, (1997). On the Calibration of Probability Judgments: Some Critical Comments and Alternative Perspectives. *Journal of Behavioral Decision Making, 10,* 269-278.

Klayman J, Soll J., González-Vallejo C., & Barlas S., (1999). Overconfidence: It Depends on How, What, and Whom You Ask, *Organizational Behavior and Human Decision Processes, 79:3,* 216-247.

Koeszegi B., (2006). Ego utility, overconfidence, and task choice, *Journal of the European Economic Association, 4:4,* 673—707.

Krajc M., & Ortmann A., (2008). Are the Unskilled Really That Unaware? An alternative explanation. *Journal of Economic Psychology,* in press.

Kruger J., & Dunning D., (1999). Unskilled and Unaware of It: How Difficulties in Recognizing One's Own incompetence Lead to Inflated Self-Assessment. *Journal of Personality and Social Psychology, 77,* 1121-1134.

Krueger I.J., & Mueller A.R. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology, 82,* 180-188.

Murphy A.H., & Winkler R.L., (1984). Probability Forecasting in Meteorology. *Journal of the American Statistical Association, 79,* 489-500.

Niederle M., & Vesterlund L., (2007). Do Women Shy Away From Competition? Do Men Compete Too Much?, *The Quarterly Journal of Economics, 122(3),* 1067-1101.

Rydval O., & Ortmann A., (2004). How financial incentives and cognitive abilities affect task performance in laboratory settings: an illustration. *Economics Letters, 85 (3)*, 315-320.

Appendix A

Table 1. *Summary statistics for actual scores and predictions in Experiment 1.*

	Midterm exam					Final exam		
	Actual score	Predictions M1		Predictions M2		Actual score	Predictions F	
		Score	Percentile	Score	Percentile		Score	Percentile
Mean	40.52	74.42	0.25	70.67	0.28	44.28	59.22	0.38
St. Dev.	27.91	11.46	0.15	19.25	0.19	24.04	21.03	0.23

Table 2. *Miscalibration in Experiment 1 (p-value for the hypothesis of the corresponding mean being equal zero in parentheses).*

	M1		M2		F	
	Score	Percentile	Score	Percentile	Score	Percentile
Mean overestimation	33.41 (0.000)	0.23 (0.000)	29.22 (0.000)	0.20 (0.000)	14.28 (0.000)	0.11 (0.002)
Mean absolute deviation	36.60 (0.000)	0.29 (0.000)	31.00 (0.000)	0.26 (0.000)	18.54 (0.000)	0.18 (0.000)

Table 3. *Paired t-test (p-value) and Cohen's d (effect size) for pairwise comparisons of mean overestimation and mean absolute deviation between predictions in Experiment 1.*

		M1-M2		M1-F		M2-F	
		Score	Percentile	Score	Percentile	Score	Percentile
Mean overestimation	p-value	0.139	0.247	0.001	0.139	0.001	0.384
	Cohen's d	0.17	0.096	0.88	0.46	0.75	0.35
Mean absolute deviation	p-value	0.027	0.030	0.000	0.042	0.001	0.286
	Cohen's d	0.28	0.15	1.06	0.61	0.76	0.42

Table 4. *The estimated intercepts and slopes of linear regressions of predicted scores on real scores and predicted percentiles on real percentiles in Experiment 1 (standard errors in parentheses).*

	M1		M2		F	
	Score	Percentile	Score	Percentile	Score	Percentile
Intercept	69.45 (3.08)	0.177 (0.043)	53.40 (3.89)	0.155 (0.048)	30.64 (4.72)	0.123 (0.050)
Slope	0.121 (0.063)	0.144 (0.080)	0.417 (0.078)	0.262 (0.085)	0.636 (0.093)	0.522 (0.089)

Table 5. *Summary statistics for actual scores and estimates in Experiment 2.*

Task 1	Stage 1			Stage 2		
	Actual score	Predictions		Actual score	Predictions	
		Score	Percentile		Score	Percentile
Mean	6.85	7.53	0.34	7.70	9.16	0.35
St. Dev.	3.55	3.54	0.24	3.63	4.10	0.23

Task 2	Stage 1			Stage 2		
	Actual score	Estimates		Actual score	Estimates	
		Score	Percentile		Score	Percentile
Mean	16.41	14.13	0.26	12.71	13.10	0.33
St. Dev.	1.99	2.92	0.18	2.32	3.09	0.19

Table 6. *Miscalibration in Experiment 2 (p-value for the hypothesis of the corresponding mean being equal zero in parentheses).*

Task 1	S1		S2 overall		S2 with FB		S2 without FB	
	Score	Percentile	Score	Percentile	Score	Percentile	Score	Percentile
Mean overestimation	0.62 (0.014)	0.11 (0.015)	1.45 (0.000)	0.094 (0.030)	1.00 (0.022)	0.018 (0.575)	1.87 (0.002)	0.16 (0.034)
Mean absolute deviation	1.34 (0.000)	0.26 (0.000)	1.73 (0.000)	0.21 (0.000)	1.29 (0.002)	0.11 (0.000)	2.13 (0.000)	0.30 (0.000)

Task 2	S1		S2 overall		S2 with FB		S2 without FB	
	Score	Percentile	Score	Percentile	Score	Percentile	Score	Percentile
Mean overestimation	-2.28 (0.000)	0.15 (0.002)	0.39 (0.417)	0.11 (0.036)	-0.079 (0.905)	0.0018 (0.983)	0.73 (0.288)	0.20 (0.006)
Mean absolute deviation	2.81 (0.000)	0.26 (0.000)	2.63 (0.000)	0.30 (0.000)	2.43 (0.000)	0.28 (0.000)	2.85 (0.000)	0.32 (0.000)

Table 7. *Paired t-test (p-value) and Cohen's d (effect size) pairwise comparison of mean overestimation and mean absolute deviation between predictions in Experiment 2.*

Task 1		S1-S2 overall		S1-S2 with FB		S1-S2 without FB	
		Score	Percentile	Score	Percentile	Score	Percentile
Mean overestimation	p-value	0.054	0.304	0.184	0.481	0.176	0.470
	Cohen's d	-0.43	0.063	-0.22	0.39	-0.59	-0.16
Mean absolute deviation	p-value	0.549	0.012	0.559	0.001	0.301	0.590
	Cohen's d	-0.23	0.25	0.039	0.96	-0.44	-0.19

Task 2		S1-S2 overall		S1-S2 with FB		S1-S2 without FB	
		Score	Percentile	Score	Percentile	Score	Percentile
Overestimation	p-value	0.000	0.443	0.002	0.425	0.001	0.805
	Cohen's d	-0.91	0.097	-0.80	0.44	-0.98	-0.16
Mean absolute deviation	p-value	0.805	0.537	0.106	0.320	0.290	0.944
	Cohen's d	0.088	-0.20	0.25	-0.069	-0.020	-0.30

Table 8. *The estimated intercepts and slopes of linear regressions of predicted scores on real scores and predicted percentiles on real percentiles in Experiment 2 (standard errors in parentheses).*

Task 1	S1		S2 with feedback		S2 without feedback	
	Score	Percentile	Score	Percentile	Score	Percentile
Intercept	1.41 (0.52)	0.204 (0.064)	0.52 (1.21)	0.095 (0.063)	2.65 (1.11)	0.211 (0.093)
Slope	0.886 (0.067)	0.29 (0.12)	1.06 (0.15)	0.72 (0.14)	0.90 (0.13)	0.21 (0.16)

Task 2	S1		S2 with feedback		S2 without feedback	
	Score	Percentile	Score	Percentile	Score	Percentile
Intercept	3.04 (3.15)	0.194 (0.044)	5.76 (4.17)	0.359 (0.083)	8.17 (3.27)	0.313 (0.063)
Slope	0.68 (0.19)	0.171 (0.087)	0.56 (0.31)	0.037 (0.175)	0.40 (0.26)	-0.036 (0.112)

Appendix B

Figure 1. Predicted own scores versus real scores (left), and percentiles versus real percentiles (right) in the midterm predictions 1 and 2, and the final prediction.

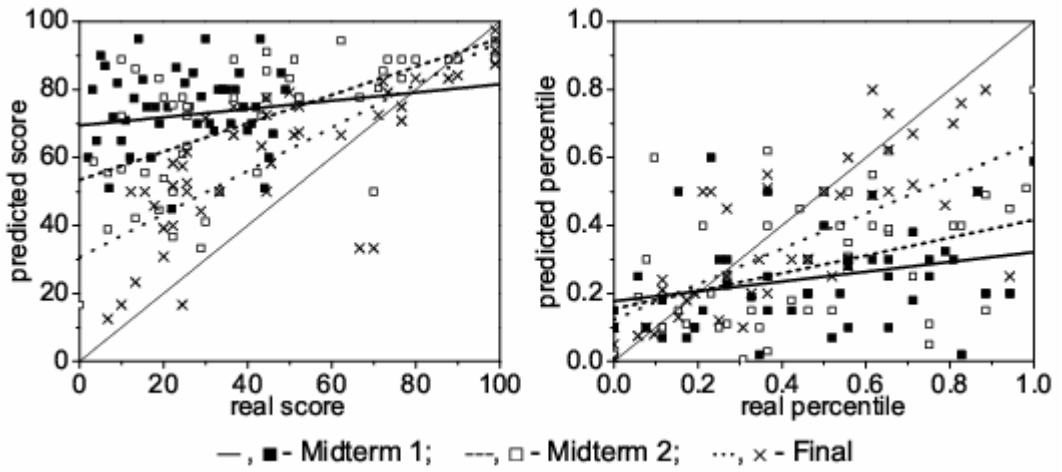


Figure 2. Mean predicted percentiles for subjects binned in performance quartiles in Experiment 1 (with lower quartiles corresponding to higher performance).

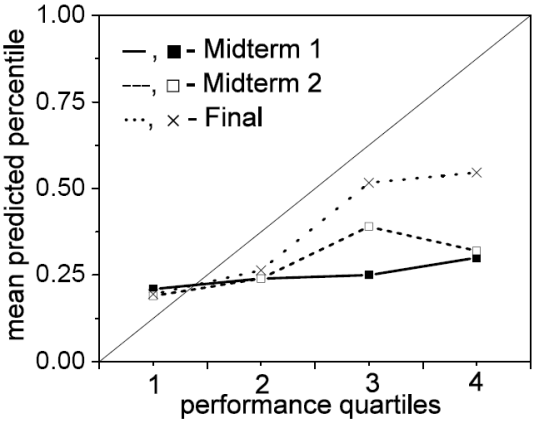
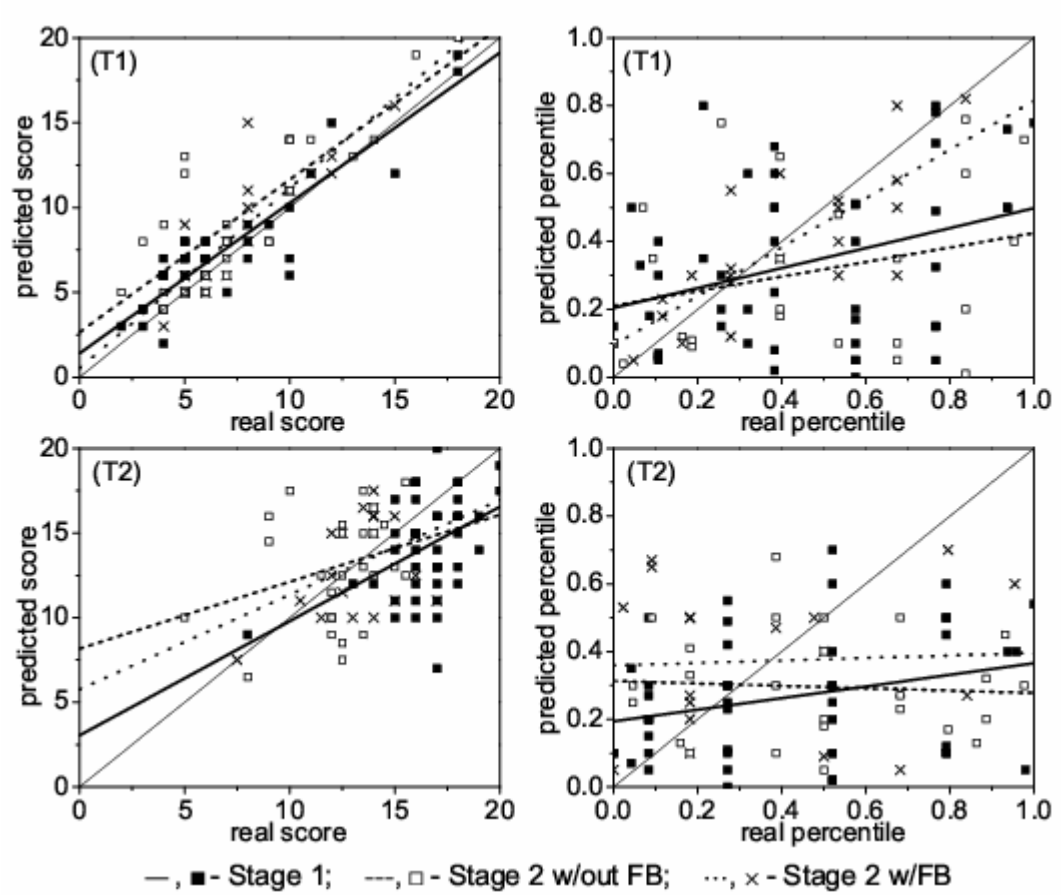


Figure 3. Predicted own scores versus real scores (left), and percentiles versus real percentiles (right) in the math skill task (T1) and general knowledge task (T2) at stage 1, stage 2 without specific feedback, and stage 2 with specific feedback.



Appendix C - the timeline and structure of the experiments

<table>
<tr><td rowspan="2"></td><td rowspan="2">Experiment 1 (field)</td><td colspan="2">Experiment 2 (lab)</td></tr>
<tr><td>Task 1 (skill)</td><td>Task 2 (general knowledge)</td></tr>
<tr><td>Week 1</td><td>Midterm predictions 1</td><td>Performance
Predictions</td><td>Performance
Predictions</td></tr>
<tr><td>Weeks 2-4</td><td>Acquiring general information</td><td colspan="2" rowspan="3">Acquiring general information</td></tr>
<tr><td>Week 5</td><td>Midterm predictions 2
Midterm performance</td></tr>
<tr><td>Weeks 6-8</td><td>Acquiring general and specific information</td></tr>
<tr><td rowspan="2">Week 9</td><td rowspan="2">Final predictions
Final performance</td><td colspan="2">Half of subjects receive specific information</td></tr>
<tr><td>Performance
Predictions</td><td>Performance
Predictions</td></tr>
</table>