Are the unskilled doomed to remain unaware?

Dmitry Ryvkin,* Marian Krajč,† and Andreas Ortmann‡

December 15, 2009

Abstract

The *unskilled-and-unaware problem* is a cognitive illusion resulting in a negative relationship between one's skill level and self-assessment bias: the less skilled are, on average, more unaware of the absolute and relative quality of their performance. In this paper, we study whether, and to what extent, the miscalibration (largely, overconfidence) of the unskilled can be reduced by feedback. We report the results of two studies, one in a natural setting and one in a more controlled setting, where subjects make incentivized judgments of their absolute and relative performance in various tasks and feedback conditions. In the first study, subjects improve their calibration after being exposed to naturally available information in the form of environmental feedback (i.e., feedback about the nature of the task) and calibration feedback (i.e., feedback about one's absolute and relative performance), but it is impossible to separate the effects of the two types of feedback. In the more controlled setting of the second study, we identified a positive effect of calibration feedback alone. In both studies, it is the unskilled who improve their calibration most. Our results suggest that the unskilled may not be doomed to be especially unaware. We also identify an important difference between the effects of feedback on the calibration of absolute and relative performance judgments. While the calibration of absolute performance judgments is more uniformly amenable to feedback, there appears to be a residual miscalibration of relative performance judgments by the unskilled that we attribute to self-image.

Keywords: calibration, judgment errors, unskilled, unaware, metacognition, self-image, experiment

*Corresponding author: Dmitry Ryvkin, Department of Economics, Florida State University, Tallahassee, FL 32306-2180; e-mail: dryvkin@fsu.edu; tel.: +1-850-878-2679, fax: +1-850-644-4535.

†The Czech National Bank, Prague, the Czech Republic; e-mail: Marian.Krajc@cnb.cz.

‡Department of Economics, Australian School of Business, The University of New South Wales, Sydney, Australia; e-mail: a.ortmann@unsw.edu.au or aortman@yahoo.com; tel.: +61-2-9385-3345.

1 Introduction

The so-called *unskilled-and-unaware problem* was first identified by Kruger and Dunning (1999). The authors conducted several experiments in which subjects were asked to estimate their absolute and relative performance in various tasks. Kruger and Dunning (1999) established three regularities: (i) poor performers (the "unskilled") overestimated their absolute and relative performance; (ii) top performers (the "skilled") underestimated their absolute and relative performance; and (iii) these miscalibrations were typically highly asymmetric in that many more unskilled overestimated their performance than the skilled underestimated theirs; often the unskilled did so quite dramatically.

Similar patterns of miscalibration were found in a number of studies across several domains (for reviews, see, e.g., Dunning 2005; Ehrlinger, Johnson, Banner, Kruger, and Dunning 2008). The original explanation for the phenomenon, provided by Kruger and Dunning (1999), is that the unskilled also lack the metacognitive ability to realize their incompetence and are thus afflicted by a double curse. This explanation implies that there is a fundamental relationship between one's skill level and one's ability to assess the quality of one's own and others' performance.[1]

A number of follow-up studies criticized Kruger and Dunning (1999) and proposed alternative explanations for the observed over- and underconfidence patterns. For example, Krueger and Mueller (2002) showed that even if all subjects have equal difficulty assessing their skill levels, the unskilled-and-unaware problem may arise in the data due to the interaction of two effects: the regression-to-the-mean effect related to the unreliability of performance as a measure of skill level, and the better-than-average effect – the belief of most people that they are above average (for a review see, e.g., Alicke and Govorun 2005). Burson, Larrick, and Klayman (2006) found that the unskilled-and-unaware problem can be mitigated or even reversed (with more miscalibration exhibited by the skilled) by manipulating the perceived difficulty of the task. Krajč and Ortmann (2008) provided a statistical explanation for the unskilled-and-unaware problem. They showed that the over- and underconfidence patterns found by Kruger and Dunning (1999) can be generated by combining the assumptions of the noisy perception of one's own performance, regression to the mean, and a skewed distribution of skills in the group under consideration.

However, in a recent article, Ehrlinger et al. (2008) addressed the concerns of the critics of Kruger and Dunning (1999) and asserted that the inability of the unskilled to

[1] In our choice of terminology we follow Kruger and Dunning (1999). The terms "unskilled" and "skilled" refer to relative performance in the group under consideration and should be understood as short-hand notation for "less skilled" and "more skilled" within the group.

assess their performance is a robust cognitive phenomenon contributing the most to the unskilled-and-unaware problem. In a series of experiments in real-world settings (students assessing their performance on an exam, university debate tournament participants assessing their performance in a debate, gun owners assessing their knowledge of gun operation and safety) and in the lab, Ehrlinger et al. (2008) corrected for the measurement error and statistical unreliability proposed as an explanation for the problem by Krueger and Mueller (2002). They also addressed the critique by Burson et al. (2006) who stated that the degree of the problem depends on the task difficulty. In all of their experiments of varying task difficulty, Ehrlinger et al. (2008) found strong overconfidence in the assessment of own absolute and relative performance by the unskilled. Ehrlinger et al. (2008) also explored the role of incentives in providing accurate self-assessment responses. In three of their five experiments, they used monetary and social incentives of various magnitudes for accurate self-assessments and found no significant improvement in calibration. Finally, using meta-analysis of four studies, Ehrlinger et al. (2008) showed that it is the miscalibration in self-assessment, not the assessment of others, which contributes most to the overconfidence of the unskilled.

Ehrlinger et al. (2008) conclude that the unskilled-and-unaware problem is a persistent feature of decision making. In some of their experiments, subjects had prior experience with the tasks and undoubtedly received feedback about their performance in the past. For example, students had taken multiple exams, and debate tournament participants had gone through multiple such tournaments, with known results. Similar findings were reported in other studies: for example, Haun, Zeringue, Leach, and Foley (2000) found that medical lab personnel cannot adequately assess the accuracy of procedures they routinely perform on the job. Ehrlinger et al. (2008) express hope that "[...] future research might shed light on the motivational and cognitive contributors to this failure to update predictions in the light of negative feedback on past performances."

In this manuscript, we explore the effect of information, or feedback on past performances, on the unskilled-and-unaware problem. We report the results of two studies. The first study uses a real-world setting similar to one of those used by Ehrlinger et al. (2008) – students predicting their performance on class exams. The second study is a laboratory experiment employing two tasks – a mathematical skill task, and a general knowledge task. In both studies, subjects estimate their performance repeatedly, which allows us to measure the impact of feedback on miscalibration.

Subjects in our two studies demonstrate, initially, miscalibration patterns consistent with the unskilled-and-unaware problem. We show, however, that experience and information improve calibration, especially for the unskilled. Thus, our results suggest that the

unskilled-and-unaware problem can be reduced with feedback, and the unskilled are not doomed to be especially unaware, at least in absolute performance judgments. For relative performance judgments, we find that although the unskilled improve their calibration significantly, and more than the skilled, they remain more overconfident. We conjecture that this difference can be attributed to the negative effect of low relative placement on self-image.

In the remainder of this section we review the relevant literature on the impact of feedback on calibration and discuss the motivation for, and placement of, our study in more detail.

1.1 Prior studies on the impact of feedback on calibration

The major part of the unskilled-and-unaware problem is the strong miscalibration (overconfidence) of the unskilled. Overconfidence has been identified as a pervasive feature of decision making in numerous studies using laboratory and real-world settings across cultures (for recent reviews see, e.g., Hoffrage 2004 and Dunning 2005; see also Moore and Healy 2008 and Ehrlinger et al. 2008, and references therein; for cross-cultural analysis of overconfidence, see, e.g., Wright, Phillips, Whalley, Choo, Ng, Tan, and Wisudha 1978, Whitcomb, Önkal, Curley, and Benson 1995, Lee, Yates, Shinotsuka, Singh, Onglatco, Yen, Gupta, and Bhatnagar 1995, Yates, Lee, and Bush 1997).

Overconfidence may have negative consequences (Dunning, Heath, and Suls 2004), such as medical diagnostic errors (e.g., Smith and Dumont 1997, Haun et al. 2000) and economic losses due to excessive market entry (e.g., Camerer and Lovallo 1999) or overexposure to risk (e.g., Malmendier and Tate 2005). It is, therefore, important to understand how overconfidence, and miscalibration in general, can be reduced. Although training and experience seem the most natural ways to improve calibration, they appear to work well in some domains, such as weather forecasting (e.g., Murphy and Winkler 1984), horse race betting (e.g., Johnson and Bruce 2001), and games of skill and chance (e.g., bridge; see Keren 1997), but not in others, such as investment (see, e.g., Chen, Kim, Nofsinger, and Oliver 2007).

In experimental studies, miscalibration can be measured in a number of ways (Moore and Healy 2008). One of the most common vehicles for observing miscalibration is the probability judgment task: a person answers a number of questions, and for each question provides an estimate of the probability that the answer is correct. Typically, the average estimated probability of being correct is higher than the actual proportion of correct answers, indicating overconfidence. A number of studies addressed the issue of reducing overconfidence in probability judgments through experience (iteration) and feed-

back. Some of these attempts were successful. For example, Koriat, Lichtenstein, and Fischhoff (1980) found that overconfidence can be significantly reduced by asking subjects to list the reasons for and against each of the alternatives in dichotomous choice questions. Explicitly stating the reasons against the chosen alternative and for the alternative that ultimately was not chosen could mitigate the underweighting of the evidence against the alternative identified as a reason for overconfidence by McKenzie (1997). A related phenomenon – option fixation – was studied as a contributor to overconfidence by Sieck, Merkle, and Van Zandt (2007). The authors found that presenting the alternatives independently reduces the familiarity bias in judgments and improves calibration. Lichtenstein and Fischhoff (1980) showed that calibration can be learned through direct feedback on the aggregate accuracy of probability judgments (*calibration feedback*; see also Sieck and Arkes 2005). In their study, subjects became better calibrated very quickly but did not improve much later on; also, they could generalize their calibration learning to some tasks but not others.

However, miscalibration was found to be robust with respect to feedback in other studies. For example, Sharp, Cutler, and Penrod (1988) let subjects go through four consecutive blocks of probability judgment tasks, with half of the subjects receiving feedback about prior performance after each block, and found no significant effect of feedback on calibration. Pulford and Colman (1997) studied the effect of feedback and task difficulty on overconfidence. In their experiments, subjects went through four blocks of general knowledge questions, with three levels of difficulty and feedback/no feedback conditions. Hard questions produced overconfidence, easy questions produced underconfidence, and no significant effect of external feedback was found. The authors found, however, somewhat better calibration by all subjects in later blocks for hard questions, with intrinsic feedback through self-monitoring suggested as an explanation.

Arkes, Christensen, Lai, and Blumer (1987) showed that the perceived difficulty of the questions plays an important role in the effectiveness of feedback. In one of their experiments, one half of the subjects received training questions that seemed easy, and the other half received questions that seemed hard, whereas in reality all of the questions were hard. Half of all subjects then received feedback in the form of correct answers. The group that was "tricked" with the seemingly easy questions and received feedback exhibited underconfidence in the subsequent judgments. Additionally, Arkes et al. (1987) conducted an experiment where half of the subjects knew that they would have a group discussion of their answers after the session. Those subjects exhibited less overconfidence than the control group. Moore and Cain (2007) explored the relationship between task difficulty and overconfidence. The authors found that, with feedback, overconfidence

prevails in easy tasks, while underconfidence is observed in hard tasks. As explained by Moore and Cain (2007), subjects recognize easy tasks as such, but fail to fully take into account that they are also easy for others; similarly, subjects fail to fully take into account that hard tasks are hard for others. This so-called reference group neglect is responsible for the over- and underconfidence, depending on task difficulty. The effect is not eliminated by feedback, as subjects tend to weigh information about themselves more than information about others. The authors do not compare feedback and no feedback conditions, however, and it is impossible to say whether feedback would at least somewhat reduce the observed miscalibrations.

Stone and Opel (2000) explored how two distinct aspects of probability judgment – calibration and discrimination (resolution) – are affected by two types of feedback – environmental feedback and performance feedback. Environmental feedback is defined as the substantive information about the event to be predicted, while performance feedback is the information on one's accuracy, for example, how many questions were answered correctly, or how overconfident the person is. They found that environmental feedback improved discrimination but not calibration, and led to more overconfidence in the easy task, while performance feedback improved calibration but not discrimination. Overall, the impact of feedback was stronger for the harder task, which the authors explain simply by there being more room for improvement there. The major conclusion is that different training procedures should be employed to master calibration and discrimination that are, thus, independent cognitive skills governed by distinct underlying psychological processes.

Sieck and Arkes (2005) studied decision aid neglect – the tendency of decision makers to underuse available resources that may improve the accuracy of professional judgments. The question they addressed was whether overconfidence explains decision aid neglect, and if so, can reducing overconfidence promote decision aid use? To reduce overconfidence, the authors used two feedback conditions during or after the "training" segment of their experiment. In the first feedback condition, correct answers were provided after each judgment; this feedback condition did not improve calibration as compared to the control (no feedback) group. In the second condition, direct calibration feedback (the percentage of correct judgments, the average perceived accuracy, and an evaluation – good calibration, overconfidence, or underconfidence) was provided. Only the latter type of feedback significantly reduced overconfidence. Importantly, Sieck and Arkes (2005) distinguish between confidence in a single item and in a set of items. They conjecture that the latter should be easier to calibrate through feedback because the situation-specific cues people use to justify their judgments for each item lose their power at the aggregate level (see also Sniezek and Buckley 1991).

Multiple-item confidence tasks are typical for the previous research on the unskilled-and-unaware problem. In one of their experiments, Kruger and Dunning (1999) let subjects see the responses of their peers and revise their self-assessments. As a result, the unskilled inflated their overconfidence even more, while the skilled improved calibration by reducing underconfidence. At the same time, Kruger and Dunning (1999) showed that environmental feedback helped the unskilled improve calibration. Several authors studied students' assessments of their performance in class. Hacker, Bol, Horgan, and Rakow (2000) used predictions and postdictions of exam scores in three exams during a semester. Students received feedback about their prior results after each exam, and the relationship between self-assessment and performance was emphasized. The authors found that high-performing students improved the accuracy of both predictions and postdictions, however, low-performing students did not improve their prediction accuracy. Interestingly, it was also found that students mainly based their predictions not on their prior performance (feedback) but on their prior predictions. Thus, the better-performing students appeared more accurate as long as their performance was improving. Hacker et al. (2000) also studied students' test preparation patterns and found no effect of feedback on those. The results are explained by the attribution bias: low-performing students see others as responsible for their problems, and are not willing to update their self-assessment despite negative feedback. The authors also mention lack of incentives to provide accurate performance predictions as a possible cause for the persistence in biased self-assessments among the unskilled.[2] In a similar setting, Ferraro (2006) measured the postdiction of absolute scores and relative standings on three consecutive exams. Unlike in Hacker et al. (2000), the assessment accuracy was incentivized by nontrivial payments. All students showed improved calibration over time, but the better-performing students improved more, and thus the unskilled-and-unaware problem did not go away.

1.2 Motivation and research agenda

It is our main goal to study the impact of feedback on miscalibration, especially for the unskilled. In line with the literature on the unskilled-and-unaware problem, we focus on multiple-item judgment tasks, as miscalibration seems to be more universally susceptible to feedback in those (Sniezek and Buckley 1991, Sieck and Arkes 2005). Similar to Kruger and Dunning (1999), Ehrlinger et al. (2008), Hacker et al. (2000), and Ferraro (2006), among others, we use a natural setting of students making predictions about their performance on exams. Previous studies of the role of feedback in this context (Hacker

[2]For a study of the role of incentives in miscalibration, see, e.g., Cesarini, Sandewall, and Johannesson (2006), Hoelzl and Rustichini (2005).

et al. 2000, Ferraro 2006) found that the unskilled improve their calibration less than the skilled, if at all. This finding adds another dimension to the unskilled-and-unaware problem: not only are the unskilled unaware of their incompetence, but they are also less able (or willing) to learn about it.

In our view, there are several open questions still left to be answered regarding this result. First, as noted by the authors, students in the study of Hacker et al. (2000) did not have any incentives to provide accurate estimates of their performance. In these circumstances, factors such as the attribution bias, self-image, or the ego utility argument of Koeszegi (2006) stating that low-performing subjects may be unwilling to accept the fact that they are at the bottom of the class could have caused the persistence in the overconfidence of the unskilled. In their study of rank-dependent search behavior, Falk, Huffman, and Sunde (2006) allowed subjects to choose whether they want their true ranking to be revealed to them at the end of the experiment, and a substantial part of low-performing subjects declined. Additionally, Hacker et al. (2000) did not ask students to assess their relative performance. Ferraro (2006) collected estimates of both absolute and relative performance and used monetary incentives for accurate self-assessments, but he only analyzed postdictions, and his results for postdictions of absolute scores are similar to those of Hacker et al. (2000). Second, neither study had a control group of students who would not receive feedback. It is difficult, if not impossible, to have such a control group in a natural setting, therefore a complementary laboratory study can be an insightful alternative.

Third, both Hacker et al. (2000) and Ferraro (2006) analyzed miscalibration measured as the difference between a student's estimated and true performance on a number of different exams, and used their performance on each of those exams as a skill measure to infer the impact of feedback on the unskilled-and-unaware problem. We argue that this approach may have potential confounding effects. Suppose a student performed poorly on exam 1 and exhibited high overconfidence, as measured by the difference between the assessed and actual performance on that exam. Then, suppose the student performed better on exam 2. Her overconfidence will decrease, but this student will no longer be considered "unskilled" from the perspective of exam 2 performance. Instead, some other student who, perhaps, performed better on exam 1 and worse on exam 2, and, as a result, exhibited larger overconfidence on exam 2, will be part of the "unskilled" group for exam 2. Thus, if the "unskilled" group changes from one exam to the next, it is likely that they will continue to appear relatively unaware. However, this is not the effect we are after. The question is whether a student who performed poorly on exam 1 and showed overconfidence can learn, with feedback, to be calibrated better on exam 2. Ideally, to

isolate the effect of feedback, in this analysis we should only consider students who did not change their performance between the two exams. Our design allows for this by letting students make predictions for the same exam twice. It is impossible, however, to analyze the effect of calibration feedback in this fashion, therefore we also let students make predictions for two different exams. In the latter case, we control for the effect of the change in performance between exams on the change in miscalibration.

There are reasons to believe that the unskilled should benefit from feedback more, not less, than the skilled. First, the unskilled typically start out with much larger self-assessment biases, i.e. they have more room for improvement. Second, with feedback, the unskilled are more likely than the skilled to perceive the task as hard, which leads to lower overconfidence, at least in single-item probability judgment tasks (Arkes et al. 1987, Stone and Opel 2000).

To address these issues, we conducted two studies. In the first study, students made predictions about their absolute scores and relative standings on two exams. Importantly, the first two predictions were made at two different points in time but about the same exam, whereas the third prediction was made at yet another (later) point in time about a different exam. Although, similar to prior studies in this setting, we were unable to compare the calibration of students with and without feedback for the same task, we could perform such a comparison for the same students across time. The goal of the second study was to test the effect of calibration feedback in a controlled environment. Subjects (the same students) performed two tasks – a mathematical skill task and a general knowledge task. Both tasks were performed twice, with subjects making postdictions about their absolute and relative performance each time. The second time, half of the subjects received direct calibration feedback about their previous performance.

We proceed by presenting the design and results of each study, with a separate discussion section, and conclude with a general discussion of our findings.

2 Study 1

2.1 Subjects

The same subjects participated in both studies. Each year, CERGE-EI in Prague (in the 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, based on their results in the prep, for graduate studies. Students may anticipate that, due to the initial selection process, their peers in the prep are likely to have been among the best

in their classes at their home universities. However, when students arrive at CERGE-EI, they, with rare exceptions, meet for the first time and have minimal information about their peers' abilities.

The prep lasts for 9 weeks, during which all students typically take four courses: microeconomics, macroeconomics, mathematics, and English (academic writing). In each of the four courses, they have regular homework assignments, a midterm exam, and a final exam. For the present study, we use the data on performance and self-assessment in the microeconomics course from two cohorts of prep students: Cohort 1, of summer 2007 (58 students, 36% female, ages from 21 to 44, average age 26), and Cohort 2, of summer 2008 (53 students, 49% female, ages from 20 to 34, average age 24).

2.2 Materials, design, and procedures

Subjects were asked to predict their performance, both on an absolute and relative scale, in the microeconomics midterm and final exam. Subjects made predictions three times: twice for the midterm exam (in week 1 of the prep, at the end of a microeconomics class; and in week 5, right before the midterm), and once for the final exam (in week 9, right before the final).

Course instructors were not present at the time predictions were administered. Subjects were told that their predictions would not affect their grades and that no one but the researchers would see identifiable data. Questionnaire sheets were distributed to subjects, and instructions read out loud. Subjects answered the following self-assessment questions:
(i) "What is your prediction of your own score on the midterm [final] exam in microeconomics?"
(ii) "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?"

For each question, the subject with the best prediction was paid 500 Czech Korunas (CZK). At the time, the exchange rate was around 20 CZK for $1, and the average hourly wage was approximately 100 CZK. Thus, accurate predictions were incentivized by non-trivial payments.

2.3 Hypotheses

The first midterm prediction (M1) was administered at the time when students had very little information about the nature of the course, as well as about their own and their peers' abilities. The second midterm prediction (M2) followed five weeks of classes and homework during which students could acquire such information in a natural setting. The

feedback students received between predictions M1 and M2 was mainly *environmental*, as defined by Stone and Opel (2000), i.e. mostly related to the subject matter. Some indirect calibration feedback might also have been present. For example, students likely had prior beliefs about their performance on homework assignments, and subsequently saw the results. However, given the substantial differences in implementation between homework assignments and exams (collaboration is expected on homework assignments as opposed to exams; time constraints, problem difficulty, the scope of material covered, and grading are different), such calibration feedback is less informative than the direct calibration feedback (score and relative standing) received after the exam.

The final prediction (F) was administered at the very end of the semester. Between predictions M2 and F, students continued to receive environmental feedback and the indirect calibration feedback; they also received direct calibration feedback about their absolute and relative performance on the midterm exam.

One important feature of our design is that predictions M1 and M2 were made about *the same event* – the midterm exam. Thus, the change in calibration between predictions M1 and M2, if present, will be due only to environmental feedback (and, possibly, indirect calibration feedback), but not due to direct calibration feedback, or changes in performance and task difficulty. Predictions M2 and F, however, were made about two different exams, and the change in calibration between M2 and F can be caused by both types of feedback and other factors.

Another interesting design feature, which was not intentional, is that Cohorts 1 and 2 had different instructors, and thus, potentially, faced different presentation of the material, problem difficulty, and grading style. That Cohort 2, on average, had significantly lower midterm and final scores than Cohort 1, might have been a consequence. The difference between the two cohorts might have also stemmed from a systematic difference between the two groups of students. For example, although the demographics of the two groups are very similar, and the pool of potential candidates and selection process did not change across the two years, we found in Study 2 that, on average, Cohort 2 performed worse on a simple mathematical skill task. However, regardless of the reason for the difference, it can be interpreted as a variation in effective task difficulty between the two cohorts, which allows us to analyze, in a between-subjects setting, the impact of task difficulty on overconfidence, the unskilled-and-unaware problem, and feedback.

We test the following hypotheses:

Hypothesis 1.1: (a) Subjects exhibit miscalibration, mostly overconfidence; (b) overconfidence is negatively related to performance.

Hypothesis 1.2: (a) Miscalibration decreases in M2 compared to M1; (b) miscalibration

decreases in F compared to M2.
Hypothesis 1.3: Subjects with lower performance exhibit a stronger improvement in calibration, i.e., the unskilled-and-unaware problem is reduced by feedback.
Hypothesis 1.4: (a) Cohort 2 is more miscalibrated than Cohort 1; (b) Cohort 2 exhibits a stronger improvement in calibration than Cohort 1.

Hypothesis 1.1 states that subjects start out with biased self-assessments, and the unskilled-and-unaware problem is present. Hypothesis 1.2 targets the aggregate effect of feedback. It is expected that feedback improves calibration, although different types of feedback (between M1 and M2, and between M2 and F) may have different effects.

Hypothesis 1.3 describes the effect of feedback on the unskilled-and-unaware problem. Although Hacker et al. (2000) and Ferraro (2006) obtained the opposite result, we believe, as explained above, that there are strong reasons to expect that result to be reversed. Hypothsis 1.4 describes the effect of task difficulty in light of the results of Arkes et al. (1987), Moore and Cain (2007), and Stone and Opel (2000) for probability judgment tasks.

2.4 Results

For data analysis, we rescaled all exam scores to the 0-100 range; also, we transformed the responses to the relative standing estimation question into fractional percentiles, i.e. the fraction of people in the cohort whose score is below the student's score. To measure miscalibration in each of the three predictions, we use *overestimation* – the difference between the estimated and the real exam score, and the difference between the estimated and the real percentile. Across the two cohorts, students scored between 0 and 98.9 points on the midterm exam, with an average score of 33.1 and standard deviation 25.4; and between 0 and 95.8 points on the final exam with an average score of 37.0 and standard deviation 25.7.

2.4.1 Aggregate miscalibration

Row "All" in Table 1 shows the average overestimation of scores and percentiles for each prediction, with standard errors in parentheses. Overestimation is everywhere positive and highly statistically significant, i.e., on average, subjects exhibit strong overconfidence. The result supports part (a) of Hypothesis 1.1.

As seen from Table 1 (row "All"), average overestimation decreases from M1 to M2 to F. We calculated the change in the overestimation of scores and percentiles between predictions M1 and M2, and between predictions M2 and F, for each subject, and then

	Scores				Percentiles			
	M1	M2	F(f)	F(m)	M1	M2	F(f)	F(m)
Q1	58.1***	45.4***	15.9***	11.7***	0.51***	0.44***	0.27***	0.19***
	(2.7)	(4.2)	(3.5)	(3.1)	(0.04)	(0.05)	(0.05)	(0.06)
Q2	51.1***	36.3***	18.0***	10.9***	0.38***	0.21***	0.10**	0.08
	(3.1)	(3.7)	(3.5)	(2.7)	(0.04)	(0.03)	(0.04)	(0.05)
Q3	34.2***	37.0***	13.1***	14.7***	0.14***	0.09*	0.05	0.09*
	(3.7)	(3.5)	(2.5)	(3.8)	(0.03)	(0.04)	(0.03)	(0.05)
Q4	12.0***	15.2***	-3.1	5.1	-0.07**	-0.06**	-0.06**	0.03
	(4.1)	(3.5)	(2.6)	(3.6)	(0.03)	(0.03)	(0.03)	(0.03)
All	38.1***	33.3***	10.2***		0.23***	0.17***	0.09***	
	(2.6)	(2.2)	(1.7)		(0.03)	(0.03)	(0.02)	

Table 1: Mean overestimation of scores and percentiles in predictions M1, M2, and F, by performance quartiles and overall (standard errors in parentheses). For prediction F, columns F(f) and F(m) describe the breakdown by f-quartiles and m-quartiles, respectively. Significance levels: *** - $p < 0.01$, ** - $p < 0.05$, * - $p < 0.1$.

found mean changes in overestimation across all subjects. The results are reported in row "All" of Table 2. The mean differences in the overestimation of scores and percentiles between M1 and M2, and between M2 and F, are positive and statistically significant. The results support parts (a) and (b) of Hypothesis 1.2.

2.4.2 Miscalibration and performance

Figure 1 and Table 1 show the average overestimation of scores and percentiles for predictions M1, M2, and F by performance quartiles. We refer to the performance quartiles based on the midterm and final exam scores as m-quartiles and f-quartiles, respectively. Overestimation in predictions M1 and M2 is shown for m-quartiles, while overestimation in prediction F is shown for both m-quartiles (column F(m)) and f-quartiles (column F(f)).

Scores. Subjects in all performance quartiles exhibit strong overconfidence in scores in predictions M1 and M2. Overconfidence is negatively related to performance, although there is no significant difference in overconfidence between the second and third quartiles in prediction M2. In prediction F, calibration is much better than in M1 and M2. The lower three quartiles still exhibit statistically significant overconfidence, but the top quartile is perfectly calibrated. Interestingly, there is no significant difference in overconfidence across the lower three quartiles regardless of whether m-quartiles or f-quartiles are used for sorting. Thus, the monotonic negative relationship between overconfidence and performance largely disappeared.

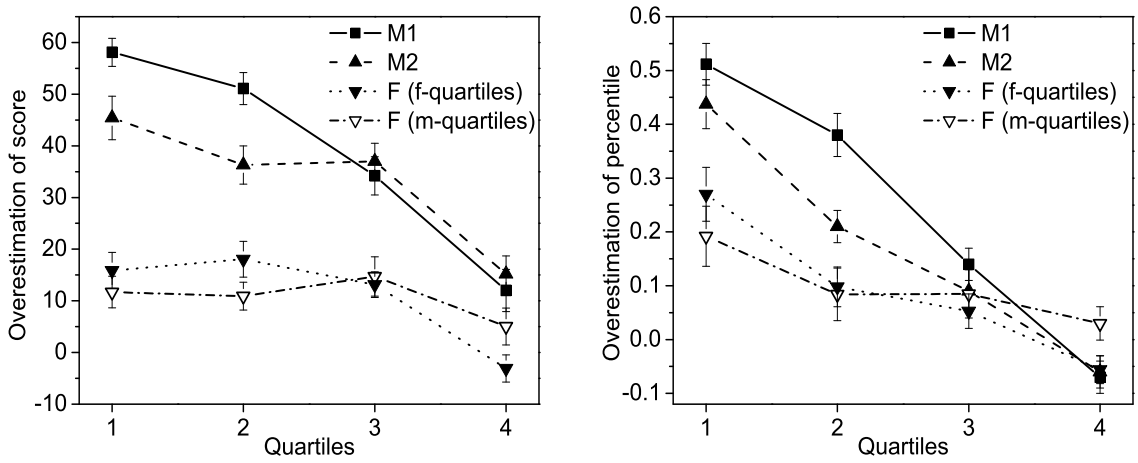

Figure 1: *Study 1*: Average overestimation of score (left) and percentile (right) by performance quartile in predictions M1, M2, and F, with error bars. The results for prediction F are shown using the breakdown by quartiles based on final performance (f-quartiles) and midterm performance (m-quartiles).

Percentiles. Subjects in the lower three quartiles exhibit statistically significant overconfidence in their relative standings in predictions M1 and M2. Subjects in the top quartile are underconfident. In prediction F, the results depend on whether m-quartiles or f-quartiles are used for sorting. Subjects in the first m-quartile are overconfident, while overconfidence is much smaller in the second through the fourth m-quartiles, and is only marginally significant in the third m-quartile. However, significant overconfidence is observed in the lower two f-quartiles, and underconfidence in the top f-quartile.

The differences between the results for m-quartiles and f-quartiles indicate the degree of "mixing," or students transitioning between performance groups between the two exams. However, as seen from Figure 1, mixing does not qualitatively affect the main result: the unskilled improve calibration more than the skilled, to the extent of eliminating the unskilled-and-unaware problem for scores, and significantly reducing it for percentiles.

2.4.3 Dynamics of miscalibration

To assess how miscalibration changed over time depending on performance, we used the difference in the overestimation of scores and percentiles between predictions M1 and M2, and between predictions M2 and F, computed for each student. Further, we calculated the average change in overestimation between M1 and M2 in each of the m-quartiles, between M2 and F in each of the m-quartiles, and between M2 and F in each of the f-quartiles. The results are presented in Table 2.

As seen from Table 2, the lower two quartiles showed a significant improvement in

	Scores			Percentiles		
	M1-M2	M2-F(m)	M2-F(f)	M1-M2	M2-F(m)	M2-F(f)
Q1	16.7**	34.9***	27.1***	0.09	0.22***	0.09
	(6.4)	(5.6)	(5.3)	(0.05)	(0.06)	(0.07)
Q2	14.8***	23.8***	19.5***	0.18***	0.12*	0.04
	(3.0)	(3.9)	(5.7)	(0.03)	(0.06)	(0.06)
Q3	-2.0	24.8***	19.7***	0.06	0.05	0.07
	(4.2)	(5.1)	(4.8)	(0.04)	(0.05)	(0.05)
Q4	-3.6	11.9**	22.5***	-0.03	-0.09**	0.03
	(3.0)	(4.8)	(4.9)	(0.03)	(0.03)	(0.04)
All	5.8**	22.3***		0.07***	0.06**	
	(2.3)	(2.6)		(0.02)	(0.03)	

Table 2: Mean difference in the overestimation of scores and percentiles between predictions, by performance quartiles and overall (standard errors in parentheses). Column M1-M2 shows the difference in overestimation between predictions M1 and M2. Columns M2-F(m) and M2-F(f) show the difference in overestimation between predictions M2 and F with the breakdown by m-quartiles and f-quartiles, respectively. Significance levels: *** - $p < 0.01$, ** - $p < 0.05$, * - $p < 0.1$.

calibration for scores between predictions M1 and M2, while the top two quartiles showed an insignificant change in calibration. At the same time, all quartiles showed a significant decrease in the overestimation of scores between predictions M2 and F. The latter result holds for m-quartiles as well as for f-quartiles. Students in the lowest quartile decrease their overestimation most.

For percentiles, a significant change in miscalibration between predictions M1 and M2 is observed only in the second quartile. Further, overestimation changes significantly between predictions M2 and F across all m-quartiles except the third one, but practically does not change in any of the f-quartiles. Students in the lowest m-quartiles improve their calibration most.

Thus, our results largely support Hypothesis 1.3, albeit with important deviations for calibration in relative standings.

2.4.4 Task difficulty

There is a significant difference in exam scores between the two cohorts of students. Cohort 1, on average, scored 15.4 higher on the midterm exam ($p = 0.002$), and 14.5 higher on the final exam ($p = 0.006$). This difference may be due to differences in instruction and

	Scores			Percentiles		
	M1	M2	F	M1	M2	F
Cohort 1 ("easier")	33.4	29.2	14.3	0.23	0.20	0.11
	(3.8)	(3.2)	(2.5)	(0.04)	(0.04)	(0.03)
Cohort 2 ("harder")	43.7	37.6	5.7	0.23	0.13	0.07
	(3.4)	(2.8)	(2.2)	(0.04)	(0.04)	(0.03)
p-value for difference	0.051	0.052	0.013	0.883	0.223	0.329

Table 3: Average overestimation of scores and percentiles in each prediction separately for two student cohorts representing different levels of effective task difficulty (standard errors in parentheses). The p-value corresponds to the null hypothesis of no difference between the two cohorts.

grading, material difficulty, or student ability.[3] Regardless of the source of the difference, however, it can be interpreted as a between-subjects variation in effective task difficulty.

Table 3 shows average overestimation of scores and percentiles for each cohort. For scores, Cohort 2 is more overconfident in predictions M1 and M2, but less overconfident in prediction F. For percentiles, there is no significant difference in overconfidence between the cohorts.

2.4.5 Regression analysis

In order to assess the effect of feedback on the unskilled-and-unaware problem at the individual level, we use linear regression analysis. One alternative can be, similar to Ferraro (2006), to treat overestimation in the three predictions for each subject as panel data and regress it on the corresponding exam scores or percentiles with dummy variables capturing the effects of feedback and exam difficulty. One problem with this approach is that the unobserved heterogeneity between subjects, such as prediction ability or risk preferences, is likely to be correlated with the explanatory variables, which makes estimation inconsistent. Another problem with using all the predictions in the same model is that such a formulation implies that feedback between predictions M1 and M2, and between predictions M2 and F, has the same effect. By our design, however, subjects were exposed to different types of feedback before and after the midterm, therefore it is more appropriate to analyze the two changes in miscalibration separately.

The approach we employ addresses both issues. For each subject i, we compute ΔOS_i^{M1M2} – the change in the overestimation of the midterm score between predictions

[3]As is evident from the results of Study 2 (discussed below), initially there was some difference in mathematical skill level between the two cohorts, which may be responsible for part of the difference in exam scores. The difference disappeared by the end of the semester, though.

M1 and M2.[4] Similar variables have been formed for the change in the overestimation of the percentile between predictions M1 and M2, and the changes in the overestimation of the score and percentile between predictions M2 and F. Further, we separately analyze the dependence of the changes in miscalibration between M1 and M2, and between M2 and F, on subjects' performance.

By design, the change in miscalibration between predictions M1 and M2 provides a clean test of the effect of feedback on miscalibration. Both predictions are made about the same exam, and the only difference between the two predictions is the feedback received by subjects by the time they made prediction M2. We estimated the following model:

$$\Delta OS_i^{M1M2} = \beta_0 + \beta_1 S_i^M + \beta_2 HARD_i + u_i. \tag{1}$$

Here S_i^M is subject i's real midterm score; $HARD_i$ is a dummy variable equal to 1 if subject i belongs to Cohort 2, and 0 otherwise; u_i is a zero-mean idiosyncratic error term; β_0, β_1, and β_2 are the unknown coefficients of interest. A similar model has been estimated for the change in the overestimation of the percentile. The results are shown in columns (1) and (2) of Table 4.

As seen from Table 4, the change in miscalibration in both scores and percentiles is negatively related to performance. Thus, the unskilled improve their calibration more than the skilled. We conclude that the unskilled-and-unaware problem is reduced as a result of environmental feedback and the indirect calibration feedback subjects received between predictions M1 and M2.

The effect of task difficulty (variable $HARD$) is significant for percentiles: the subjects who faced the harder exam improved their calibration more. For the scores, however, the effect of task difficulty is not statistically significant.

For the change in the miscalibration of scores between predictions M2 and F, we estimated the following model:[5]

$$\Delta OS_i^{M2F} = \beta_0 + \beta_1 S_i^M + \beta_2 \Delta S_i^{FM} + \beta_3 HARD_i + u_i. \tag{2}$$

[4] ΔOS_i^{M1M2} is defined as $\Delta OS_i^{M1M2} = OS_i^{M1} - OS_i^{M2}$, where OS_i^{M1} and OS_i^{M2} are subject i's overestimation of scores in predictions M1 and M2, respectively. Thus, positive ΔOS_i^{M1M2} implies a *decrease* in overestimation. We also note that ΔOS_i^{M1M2} does not contain the actual midterm score.

[5] This specification is similar to the panel data specification used by Ferraro (2006). The important advantage of our specification, however, is that we use the differenced version of the model, thereby filtering out the unobserved individual heterogeneity. Ferraro (2006) uses the random effects panel data estimator (see, e.g., Wooldridge 2002), which is consistent only under the assumption that the unobserved effect is uncorrelated with the explanatory variables (e.g., midterm and final scores). We find this assumption too restrictive.

	M1-M2		M2-F	
	Scores	Percentiles	Scores	Percentiles
	(1)	(2)	(3)	(4)
Midterm score	-0.36***		-0.27***	
	(0.09)		(0.08)	
Midterm percentile		-0.20***		-0.27***
		(0.07)		(0.08)
F-M change in score			0.56***	
			(0.14)	
F-M change in percentile				0.75***
				(0.12)
HARD	-1.84	0.085**	11.4***	-0.03
	(4.5)	(0.040)	(4.2)	(0.04)
Intercept	19.3***	0.14***	25.5***	0.24***
	(4.8)	(0.05)	(4.7)	(0.05)
N	78	80	81	83
R^2	0.18	0.15	0.47	0.46

Table 4: Estimation results for the change in the overestimation of the scores and percentiles between predictions M1 and M2, and between predictions M2 and F (standard errors in parentheses). The results are obtained by OLS estimation of Eq. (1) and a similar equation for percentiles for columns (1) and (2), and of Eq. (2) and a similar equation for percentiles for columns (3) and (4). Significance levels: *** - $p < 0.01$, ** - $p < 0.05$, * - $p < 0.1$.

Here, in addition to subject i's real midterm score, S_i^M, we control for the change in her performance between the midterm and final, $\Delta S_i^{FM} = S_i^F - S_i^M$, where S_i^F is subject i's actual final score. A similar model was estimated for percentiles. The results are presented in columns (3) and (4) of Table 4.

As seen from Table 4, the coefficient estimates on midterm scores and percentiles are negative and statistically significant in the corresponding regressions. Thus, conditional on the change in performance between the exams, the unskilled improve their calibration more than the skilled. As expected, the coefficient estimates on the change in performance are positive, indicating that improvement in a student's performance is a separate channel for improving calibration. Interestingly, unlike between M1 and M2, the effect of task difficulty is here significant for scores but not for percentiles.

2.5 Discussion

The goal of this study was to explore the effect of feedback on the unskilled-and-unaware problem in a natural setting. Recall that the unskilled-and-unaware problem is defined

as a negative relationship between one's degree of miscalibration (overconfidence) and performance. In our study, students made the first predictions of their midterm exam scores and percentiles (M1) at the very beginning of the semester. Subjects exhibited significant overconfidence in both scores and percentiles, and the degree of overconfidence was decreasing in performance, i.e. the unskilled-and-unaware problem was present. By the time of the second prediction for the same midterm exam (M2), four weeks into the semester, subjects interacted formally and informally, and received environmental feedback about the subject matter as well as indirect calibration feedback through homework assignments. Soon after the midterm, subjects received direct calibration feedback about their performance, and continued to receive environmental and indirect calibration feedback for four more weeks until the final prediction (F).

We found strong support of our hypothesis that feedback improves calibration, especially for the unskilled. For scores, there is no difference in miscalibration between the lower three quartiles of students by prediction F, and the degree of overconfidence is much lower than in prediction M1. For percentiles, the overconfidence of students in the first quartile in prediction F is still stronger than for all other students, but it is significantly lower than in prediction M1. The lower two quartiles experienced a much stronger improvement in calibration of percentiles than the upper two quartiles.

The aggregate results are confirmed by regression analysis at the individual level. We found a strong negative relationship between the *change* in overconfidence between consecutive predictions and performance, both for scores and percentiles. We also accounted for the difference between two cohorts of students (effectively, task difficulty) and found that the task difficulty has a statistically significant positive effect on the change in overconfidence in percentiles between predictions M1 and M2, and in scores between predictions M2 and F. In what follows we first discuss our main result – that, contrary to some of the prior findings, the unskilled-and-unaware problem is mitigated by feedback. Second, we discuss the role of task difficulty. Third, we discuss the limitations of this study and the reasons why a complementary laboratory experiment is necessary to clarify some of the issues.

2.5.1 Feedback and the unskilled-and-unaware problem

Between predictions M1 and M2, our subjects experienced environmental feedback (Stone and Opel 2000) in the form of lectures and homework assignments. By design, they could not have received direct calibration feedback because both predictions were made about the same event that followed prediction M2. For probability judgment tasks, the impact of environmental feedback on calibration was studied by a number of authors, with dif-

ferent outcomes. For example, Stone and Opel (2000) found that environmental feedback increased overconfidence, while Lichtenstein and Fischhoff (1977), and also Kruger and Dunning (1999), obtained the opposite result. As suggested by Stone and Opel (2000), the difference can be explained by considering the exact nature of feedback subjects received. While in their study subjects believed that they learned a lot from the information provided, in the studies of Lichtenstein and Fischhoff (1977) and Kruger and Dunning (1999) subjects had to extract and classify information on their own and thus were likely to be less confident in their acquired knowledge. In our study, the nature of environmental feedback was mostly of the latter type. Thus, insofar as the results from other types of tasks can be transferred to our setting, the improvement in calibration between M1 and M2 can be explained by the ambiguity of feedback and perceived difficulty of the tasks our subjects were receiving.

It is also possible that, although direct calibration feedback (performance feedback, Stone and Opel 2000) was not present, students could experience indirect calibration feedback by updating their beliefs about their absolute and relative standings through communication in study groups and performance feedback on homework assignments. For probability judgment tasks, Stone and Opel (2000) and Sieck and Arkes (2005) found that calibration feedback reduces overconfidence. As conjectured by Sieck and Arkes (2005), the effect should be even more pronounced in multiple-item judgment tasks.

Between predictions M2 and F, subjects received direct calibration feedback and continued to receive environmental feedback. The improvement in calibration between M2 and F is stronger than between M1 and M2. This finding is in line with the results of prior studies (e.g., Stone and Opel 2000, Sieck and Arkes 2005) that showed the importance of calibration (performance) feedback for calibration training in probability judgment tasks.

The unskilled in our study improved their calibration significantly more than the skilled. Interestingly, the unskilled-and-unaware problem practically goes away for scores, while there is some residual overconfidence of the most unskilled subjects in percentiles. As seen from Figure 1, although subjects in the first quartile exhibit the strongest improvement in calibration by prediction F, they, on average, still place themselves into the second quartile. This residual persistence in overconfidence can be explained by the self-image and ego utility arguments of Koeszegi (2006): students are not willing to admit to others and themselves that they are at the bottom of the skill distribution. Our findings suggest, however, that this cannot serve as an explanation of the results of Hacker et al. (2000), who did not find an improvement in the calibration of scores for low-performing subjects. Low scores do not have the same negative meaning for self-image as low percentiles. Even if a student believes that his score will be low, he may convince himself

that other students' scores will be low as well, thereby improving his calibration in scores without affecting self-image at the expense of staying miscalibrated in percentiles.

2.5.2 The role of task difficulty

Using the naturally occurred difference in performance between the two cohorts of students participating in the study, we analyzed the role of task difficulty in the effect of feedback on miscalibration. At the aggregate level, we compared overconfidence in scores and percentiles between the two cohorts in each prediction. For scores, we found that Cohort 1 (the cohort for which the tasks were easier) was calibrated better than Cohort 2 in predictions M1 and M2, but worse than Cohort 2 in prediction F. For percentiles, we found no significant difference in miscalibration between the two cohorts at the aggregate level.

For probability judgment tasks, Arkes et al. (1987) found that the perceived difficulty of the task is an important predictor of overconfidence. In their study, less overconfidence was observed for the task perceived as hard (although both tasks were of the same difficulty). Moore and Cain (2007) manipulated task difficulty and found that subjects are mainly overconfident for easy tasks and underconfident for hard tasks, i.e. overconfidence decreases in the actual task difficulty as well. In contrast, our subjects, at least initially, exhibit more overconfidence in scores for the harder task. However, overconfidence decreased with feedback, especially between M2 and F when direct calibration feedback was provided, and by prediction F, Cohort 2 was calibrated better than Cohort 1. For percentiles, the improvement in calibration with feedback occurred mainly between predictions M2 and F for Cohort 1, and mainly between predictions M1 and M2 for Cohort 2. Thus, it appears from the aggregate results that Cohort 1 responded mainly to direct calibration feedback, while Cohort 2 responded mainly to environmental feedback.

Our findings for scores are consistent with those of Arkes et al. (1987) and Moore and Cain (2007) in that the subjects who received lower scores on the midterm exam (Cohort 2) became less overconfident in their final prediction. The results for percentiles are consistent with the reference group neglect explanation of Moore and Cain (2007): environmental feedback between M1 and M2 made Cohort 2 subjects perceive the task as hard and reduced their overconfidence in percentiles. Having received direct calibration feedback, they reduced overconfidence even further, but not as much because they were already relatively well-calibrated. Cohort 1 subjects, on the other hand, did not perceive the task to be as difficult until they received direct calibration feedback, and hence their improvement in calibration mainly occurred between predictions M2 and F.

Although aggregate analysis of the impact of task difficulty on calibration is insightful,

our main goal is to study the relationship between miscalibration, feedback and performance. From this perspective, it is of interest to assess the role of task difficulty and feedback at the individual level conditional on performance. We found that, conditional on performance, effective task difficulty is associated with a stronger improvement in calibration of percentiles, but not scores, between predictions M1 and M2, and a stronger improvement in calibration of scores, but not percentiles, between predictions M2 and F. The former result can be explained by the interaction between environmental feedback and reference group neglect, while the latter is due to performance feedback. Indeed, between predictions M1 and M2 subjects did not see their scores, so there was no reason for the two cohorts to update their predictions of absolute scores differently, whereas reference group neglect could have a differential effect on updating their beliefs about percentiles. After the midterm, when actual scores were revealed, the differential updating of beliefs about scores became possible. At the same time, there was no longer a factor to cause between-cohort differences in updating the beliefs about percentiles.

2.5.3 Limitations and motivation for Study 2

In conducting Study 1, we faced the natural limitations of the real-world setting we used. One important limitation, in our view, was the lack of control over stimuli materials. The course instructors selected their own exam problems and used discretionary grading schemes, therefore the absolute score scale was somewhat arbitrary. For example, a student who received a score of 60 (out of 100) did not necessarily solve 60% of the exam correctly. Due to the complicated and convoluted nature of microeconomics exam problems, it is difficult to objectively quantify the accuracy of partially correct answers. This is in contrast with probability judgment tasks, in which, regardless of the complexity of the subject matter, there is no ambiguity as to the proportion of correct answers.

Another important limitation is the lack of control over feedback. Between predictions M1 and M2, and M2 and F, students received environmental feedback in the form of lectures, discussions, readings, and homework assignments, but all these activities were voluntary for students to participate in, therefore different students received different amounts of environmental feedback; moreover, these differences were likely nonrandom. Additionally, students received indirect calibration feedback on their relative standings. Here, too, the amount and accuracy of such feedback depended on how actively students sought it. As discussed by Falk et al. (2006), less skilled students were also less likely to try to find out about their standings. Finally, all students received direct calibration feedback about their absolute and relative performance on the midterm exam, therefore we did not have a control group to assess the impact of such feedback separately from

other types of feedback between predictions M2 and F.

To address these limitations, we conducted Study 2. The purpose of Study 2 was to separately explore the impact of direct calibration feedback on the unskilled-and-unaware problem in a controlled environment.

3 Study 2

3.1 Subjects, materials, design, and procedures

Subjects in Study 2 were the same two cohorts of prep students as in Study 1. Subjects were asked to postdict their performance, both on an absolute scale (score) and relative scale (percentile), in two real-effort tasks. Subjects performed the tasks and made postdictions in two stages, to which we will refer as Stage 1 (conducted around the same time as prediction M1 in Study 1) and Stage 2 (conducted around the same time as prediction F in Study 1).

The experimental sessions were administered with paper and pencil at the end of two microeconomics classes. The instructors were not present during the sessions. In each session, subjects started by performing an individual real-effort task (Task 1), then made postdictions regarding their absolute and relative performance in Task 1, then performed another real-effort task (Task 2), and finally made postdictions regarding their performance in Task 2. Instructions for each part were provided separately at the beginning of that part. Subjects did not know the sequence of events in advance; also, upon completion of Stage 1 they did not receive any feedback on their performance and did not know that Stage 2 will follow.

3.1.1 Task 1

Task 1 was a mathematical skill-oriented task. Subjects had to sum, within a 3-minute time limit, sets of five 2-digit numbers without the use of calculators (see, e.g., Niederle and Vesterlund 2007, Brueggen and Strobel 2007). We distributed sheets with 22 random summation problems. Subjects were paid 5 CZK for each correctly solved problem at Stage 1, and 10 CZK for each correctly solved problem at Stage 2.

3.1.2 Task 2

Subjects had to answer, within a 2-minute time limit, a quiz containing two-alternative geography questions (such general-knowledge tasks are widely investigated in psychology; for a review, see, e.g., Juslin, Winman, and Olsson 2000). At Stage 1, we asked for a

comparison of the population of 20 random pairs of European Union countries (*"Which of the following two countries has a larger population?"*). At Stage 2, to avoid repetition, we asked for a comparison of the population of 40 random pairs of the 50 most populated countries in the world. Subjects were paid 5 CZK for each correct comparison in each stage.

3.1.3 Performance postdictions

Upon the completion of Task 1, subjects were asked the following questions:
(i) "How many summing problems do you think you solved correctly?"
(ii) "What do you think is the percentage of people in the group who performed better than you?"

Similar questions were asked after Task 2. Subjects providing the most accurate estimates for each of these questions were paid 500 CZK.

3.1.4 Direct calibration feedback

In Stage 2, half of the subjects 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 selected randomly in a stratified manner so that the number of subjects receiving feedback was approximately equalized across Stage 1 performance quartiles.

3.1.5 Remarks

Because available time gets scarcer towards the end of the semester we decided to double the incentives at Stage 2 as compared to Stage 1, to minimize attrition. In Task 1, we paid subjects 10 CZK instead of 5 CZK for each correctly solved problem; in Task 2, we gave subjects 40 questions instead of 20 keeping the pay rate at 5 CZK per correctly answered question. The results of Rydval and Ortmann (2004), Cesarini et al. (2006), and comparison between the results of Hacker et al. (2000) and Ferraro (2006), suggest that this increase should not matter in a significant way.

In Study 2, unlike in Study 1 where the exams' content and grading scheme were selected by the instructor of the class, we used tasks that allowed us to better control for the representativeness of stimuli, which, as suggested by previous research, mitigates miscalibration (see, e.g., Gigerenzer, Hoffrage, and Kleinboelting 1991, Dhami, Hertwig, and Hoffrage 2004 and Juslin, Winman, and Olsson 2000). First, we clearly specified the reference classes of questions for each task – 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.

In Task 2, we changed the reference class at Stage 2, as compared to Stage 1, to the 50 most populated world countries instead of the countries of the European Union. We did this to avoid too many subjects answering all questions correctly, as some subjects might have, out of curiosity, learnt about the population of the EU countries after Stage 1.

Incentives play an important role in various types of studies (see, e.g., Camerer and Hogarth 1999 and Rydval and Ortmann 2004). In Study 2, we used tasks that are responsive to higher effort (e.g., for general knowledge, employing more cues, as suggested by Gigerenzer et al. 1991) and therefore we expect that monetary incentives will increase the accuracy of the given answers and thus also the measured ability. To motivate the subjects to give as precise answers as possible, we used a linear incentive scheme. Because of the number of participants, we had to make a choice between using two feedback treatments or two incentive treatments. The evidence in Cesarini et al. (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.

In Study 2, unlike in Study 1, we used performance postdictions. The nature of the tasks in Study 2 was such that it was very easy for subjects to manipulate their scores and give accurate predictions by intentionally underperforming. Although such behavior is possible also with postdictions, it is less likely because, while performing the tasks, subjects could not know with certainty that postdictions will follow. We did not see any evidence of score manipulation in the data. For better comparability, an alternative could be to administer postdictions also in Study 1, but it was institutionally unfeasible. We have chosen the design as a compromise, given that the main focus of our study is the impact of feedback on the unskilled-and-unaware problem, and we expect the impact to be qualitatively similar for predictions and postdictions, especially for percentiles.

3.2 Hypotheses

Hypotheses 2.1 through 2.4 below are similar to hypotheses 1.1 through 1.4 of Study 1. Hypothesis 2.5 describes the expected effect of direct calibration feedback as compared to the control group.

Hypothesis 2.1: (a) Subjects exhibit miscalibration, mostly overconfidence; (b) overconfidence is negatively related to performance.

Hypothesis 2.2: Miscalibration decreases in Stage 2 compared to Stage 1.

Hypothesis 2.3: Subjects with lower performance exhibit a stronger improvement in calibration.
Hypothesis 2.4: (a) Cohort 2 is more miscalibrated than Cohort 1; (b) Cohort 2 exhibits a stronger improvement in calibration than Cohort 1.
Hypothesis 2.5: Subjects who received direct calibration feedback exhibit a stronger improvement in calibration.

3.3 Results

We rescaled all scores in both tasks to the 0-20 range, and transformed the relative standing postdictions into percentile fractions, similar to the analysis in Study 1. Miscalibration is measured as the difference between the postdicted and actual scores, and between the postdicted and actual percentiles.

In Task 1 at Stage 1, subjects correctly solved between 0 and 18 problems, with an average of 6.3 and standard deviation 3.5; at Stage 2, they solved between 1 and 18 problems, with an average of 7.2 and standard deviation 3.7. The difference in performance between Stage 1 and Stage 2 is statistically significant ($p = 0.002$). In Task 2 at Stage 1, subjects correctly answered between 8 and 20 questions, with an average of 16.5 and standard deviation 2.2. At Stage 2, they answered between 5 and 17.5 (rescaled) questions, with an average of 13.0 and standard deviation 2.1. The difference between Stage 1 and Stage 2 is statistically significant ($p = 0.000$).

Task 2 turned out to be very easy for our subjects, especially at Stage 1 where about 20% of subjects answered at least 90% of the questions correctly. This has consequences for the results below, which we discuss in due course.

3.3.1 Miscalibration and performance

Table 5 shows the average overestimation of scores and percentiles in Task 1 at Stages 1 and 2, by performance quartiles and overall. Similar to Study 1, we distinguish between S1-quartiles and S2-quartiles depending on which stage's performance is the basis for grouping. Table 6 shows the same results for Task 2.

The results presented in Tables 5 and 6 are visualized in Figure 2. In Task 1, the downward-sloping dependence of overestimation on performance at Stage 1 (the solid lines) indicates the initial presence of the unskilled-and-unaware problem both for scores and percentiles. Subjects are overconfident, with the exception of the top quartile, which is perfectly calibrated in scores and underconfident in percentiles. At Stage 2, however, there is a significant difference in miscalibration patterns between S1-quartiles and S2-

	Scores			Percentiles		
	Stage 1	Stage 2 (S1)	Stage 2 (S2)	Stage 1	Stage 2 (S1)	Stage 2 (S2)
Q1	1.45***	0.71*	1.80***	0.32***	0.15**	0.39***
	(0.37)	(0.36)	(0.55)	(0.05)	(0.06)	(0.06)
Q2	1.18***	1.27*	1.67***	0.30***	0.05	0.23***
	(0.31)	(0.66)	(0.55)	(0.05)	(0.07)	(0.05)
Q3	0.40*	1.38**	1.05**	0.05	0.11	-0.02
	(0.22)	(0.58)	(0.44)	(0.05)	(0.06)	(0.04)
Q4	0.07	1.39***	0.63	-0.12***	0.10**	-0.09**
	(0.38)	(0.40)	(0.41)	(0.04)	(0.05)	(0.04)
All	0.70***	1.22***		0.12***	0.10***	
	(0.18)	(0.24)		(0.03)	(0.03)	

Table 5: *Task 1*: Mean overestimation of scores and percentiles by performance quartiles and overall (standard errors in parentheses). The results for Stage 2 are shown for the quartiles based on Stage 1 performance (S1-quartiles) and Stage 2 performance (S2-quartiles). Significance levels: *** - $p < 0.01$, ** - $p < 0.05$, * - $p < 0.1$.

	Scores			Percentiles		
	Stage 1	Stage 2 (S1)	Stage 2 (S2)	Stage 1	Stage 2 (S1)	Stage 2 (S2)
Q1	-0.67	-0.06	2.20**	0.55***	0.14	0.54***
	(0.64)	(0.75)	(1.00)	(0.04)	(0.09)	(0.05)
Q2	-0.68	0.18	-0.60	0.33***	0.13	0.33***
	(0.63)	(0.96)	(0.61)	(0.05)	(0.10)	(0.04)
Q3	-3.52***	0.38	-1.36	0.10**	0.18**	-0.02
	(0.64)	(0.80)	(1.03)	(0.04)	(0.08)	(0.06)
Q4	-2.48***	-0.24	-0.22	-0.12***	0.16***	-0.15***
	(0.42)	(0.63)	(0.51)	(0.03)	(0.05)	(0.04)
All	-1.97***	-0.006		0.19***	0.15***	
	(0.31)	(0.381)		(0.03)	(0.04)	

Table 6: *Task 2*: Mean overestimation of scores and percentiles by performance quartiles and overall (standard errors in parentheses). The results for Stage 2 are shown for the quartiles based on Stage 1 performance (S1-quartiles) and Stage 2 performance (S2-quartiles). Significance levels: *** - $p < 0.01$, ** - $p < 0.05$, * - $p < 0.1$.

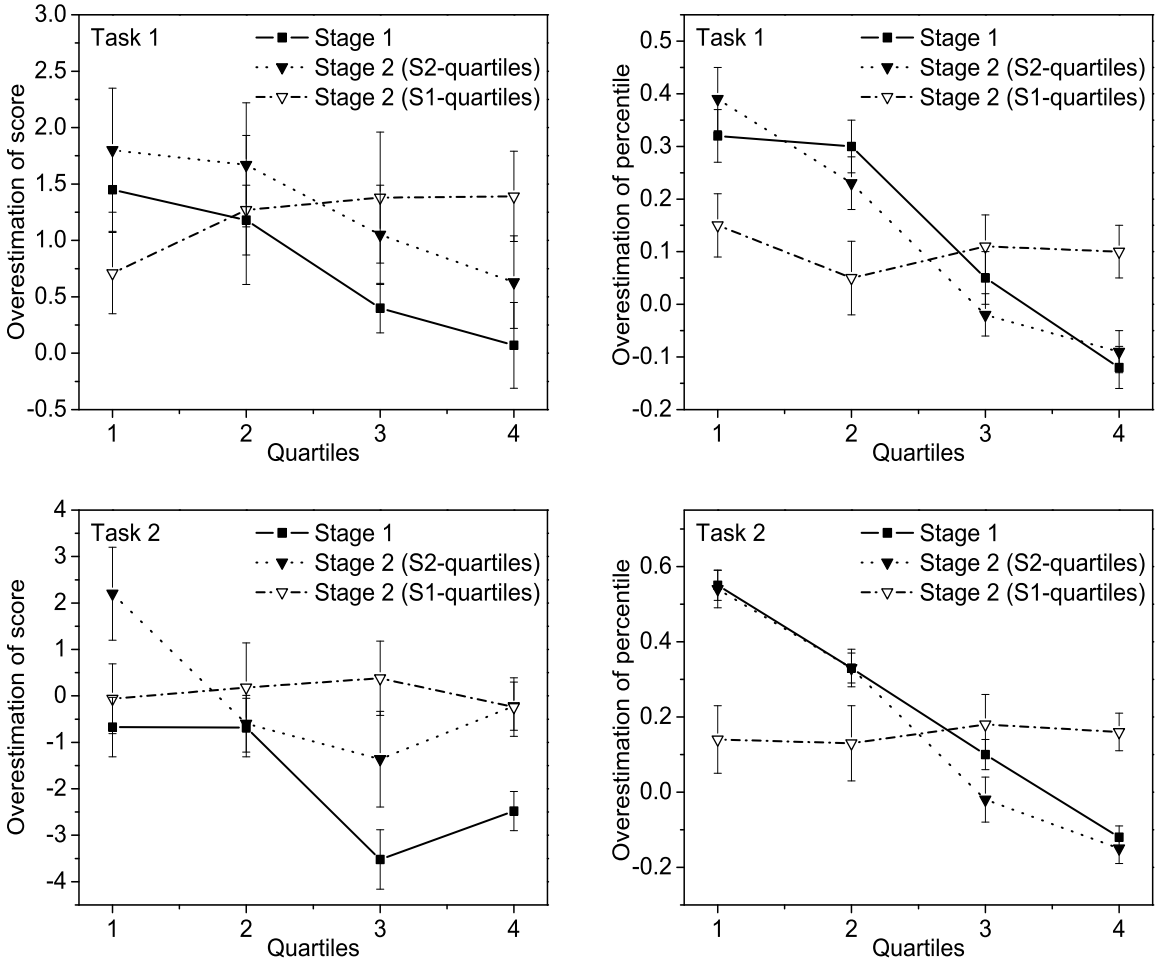

Figure 2: *Study 2*: Average overestimation of score (left) and percentile (right) by performance quartile in Task 1 (top) and Task 2 (bottom) at Stages 1 and 2, with error bars. The results for Stage 2 are shown using the breakdown by quartiles based on Stage 2 performance (S2-quartiles) and Stage 1 performance (S1-quartiles).

quartiles. For S1-quartiles, the negative dependence of miscalibration on performance goes away: those subjects who performed poorly at Stage 1 improved their calibration by becoming less overconfident, while the top two quartiles became more overconfident (the dash-dotted lines). For S2-quartiles, on the other hand, the downward-sloping dependence is preserved at Stage 2 (the dotted lines). The difference, as discussed previously, is due to the confounding effect of changes in performance between stages.

In Task 2, the pattern of initial miscalibration in scores is different. While subjects in the two lower quartiles are perfectly calibrated, the top two quartiles exhibit strong underconfidence. Thus, we observe a reversal of the unskilled-and-unaware problem, which can be attributed to the easiness of the task. At Stage 2, there is a difference between S1-quartiles and S2-quartiles similar to the one observed for Task 1. For S1-quartiles, calibration is nearly perfect across the board, whereas significant overconfidence arises in the bottom S2-quartile. For percentiles, there is no qualitative difference between the two tasks. For S1-quartiles, the unskilled start out strongly overconfident but improve their calibration considerably by Stage 2, whereas the skilled are underconfident at Stage 1, but ultimately reach the same levels of overconfidence as the unskilled. For S2-quartiles, there is no change in miscalibration.

Tables 7 and 8 show the results of statistical tests for the difference in miscalibration between Stage 1 and Stage 2 by quartile. Similar to the analysis in Study 1, we computed the difference in overestimation for each subject; thus, the statistical significance of the entries in Tables 7 and 8 corresponds to the paired t-test of the null hypothesis of no difference in miscalibration between the two stages, with individual heterogeneity taken into account.[6]

As seen from Table 7, the only significant change in the miscalibration of scores in Task 1 is in the increase in overconfidence in the top quartile. For percentiles, however, the results are consistent with Hypothesis 2.3: the unskilled improve their calibration most. In Task 2 (Table 8), the skilled overcome their underconfidence in scores, whereas for percentiles it is again the unskilled who achieve large and significant improvements.

3.3.2 Regression analysis

Similar to Study 1, we can assess the effect of feedback and incidental task difficulty on the unskilled-and-unaware problem at the individual level using regression analysis. For

[6]It would be incorrect to simply run a t-test for the difference in means between the entries in Tables 5 and 6 because they contain repeated observations from the same subjects and thus are not independent.

	Scores		Percentiles	
	S1-S2 (S1)	S1-S2 (S2)	S1-S2 (S1)	S1-S2 (S2)
Q1	0.81	-0.50	0.14*	-0.07
	(0.55)	(1.07)	(0.08)	(0.11)
Q2	0.00	-0.78	0.28**	-0.05
	(0.64)	(0.72)	(0.10)	(0.09)
Q3	-0.77	-0.25	-0.04	0.09
	(0.61)	(0.43)	(0.07)	(0.08)
Q4	-2.04***	-1.00*	-0.20***	0.05
	(0.47)	(0.53)	(0.05)	(0.07)
All	-0.66**		0.02	
	(0.31)		(0.04)	

Table 7: *Task 1*: Mean difference in the overestimation of scores and percentiles between Stage 1 and Stage 2 by performance quartiles and overall (standard errors in parentheses). The results are shown for the quartiles based on Stage 1 performance (S1-quartiles) and Stage 2 performance (S2-quartiles). Significance levels: *** - $p < 0.01$, ** - $p < 0.05$, * - $p < 0.1$.

	Scores		Percentiles	
	S1-S2 (S1)	S1-S2 (S2)	S1-S2 (S1)	S1-S2 (S2)
Q1	-0.76	-4.21***	0.40***	-0.17**
	(0.94)	(0.76)	(0.11)	(0.06)
Q2	-1.24	-1.97***	0.17*	-0.09
	(0.71)	(0.65)	(0.08)	(0.10)
Q3	-3.38***	-0.04	-0.04	0.31**
	(0.87)	(1.39)	(0.08)	(0.12)
Q4	-3.11***	-2.19***	-0.20***	0.19**
	(0.67)	(0.58)	(0.05)	(0.07)
All	-2.17***		0.07	
	(0.41)		(0.05)	

Table 8: *Task 2*: Mean difference in the overestimation of scores and percentiles between Stage 1 and Stage 2 by performance quartiles and overall (standard errors in parentheses). The results are shown for the quartiles based on Stage 1 performance (S1-quartiles) and Stage 2 performance (S2-quartiles). Significance levels: *** - $p < 0.01$, ** - $p < 0.05$, * - $p < 0.1$.

scores, we estimated the following model:

$$\Delta OS_i^{S1S2} = \beta_0 + \beta_1 S_i^{S1} + \beta_2 \Delta S_i^{S2S1} + \beta_3 HARD_i + \beta_4 FB_i + \beta_5 S_i^{S1} \cdot FB_i + \beta_6 \Delta S_i^{S2S1} \cdot FB_i + u_i. \quad (3)$$

Here, ΔOS_i^{S1S2} is the change in the overestimation of scores between Stage 1 and Stage 2 for subject i; S_i^{S1} is subject i's actual score at Stage 1; $\Delta S_i^{S2S1} = S_i^{S2} - S_i^{S1}$ is the change in performance between Stage 1 and Stage 2; $HARD_i$ is a dummy variable equal 1 if subject i belongs to Cohort 2, and 0 otherwise; FB_i is a dummy variable equal 1 if subject i received direct calibration feedback, and 0 otherwise; u_i is a zero-mean idiosyncratic error term. Equation (3) was estimated by OLS separately for each task. Similar equations have also been estimated for percentiles.

The key feature of Study 2 is the presence of controlled direct calibration feedback given randomly to half of the subjects at the beginning of Stage 2. Variable FB_i controls for the effect of feedback on the change in overestimation, whereas the interaction variables $S_i^{S1} \cdot FB_i$ and $\Delta S_i^{S2S1} \cdot FB_i$ (and similar variables for percentiles) control for the possible differential effect of feedback on calibration for the skilled and and the unskilled, and the differential effect of feedback on calibration for subjects experiencing different changes in performance between stages.

Subjects in Cohort 2, on average, solved 1.25 fewer summation problems (Task 1) at Stage 1 than subjects in Cohort 1. The difference is statistically significant at the 10% level. However, there is no significant difference in performance between cohorts at Stage 2 in Task 1, and at both stages in Task 2. We control for possible differences between cohorts using dummy variable $HARD_i$ (implying that Task 1 is harder for Cohort 2 than for Cohort 1).

The results are shown in Table 9. As seen from the table, there appears to be no significant difference between the two cohorts. Also, Stage 1 scores and percentiles themselves have no effect on the change in calibration once the changes in performance are accounted for.

Feedback affects the decrease in overconfidence positively in Task 1, but has no significant effect in Task 2. Of major interest are the estimates of the coefficient on the interaction terms. In Task 1, the interaction between the Stage 1 score and feedback is negative and statistically significant, implying that the unskilled react stronger to feedback; the same is true for percentiles. These interactions are not significant in Task 2, however.

The interaction between the change in performance and feedback is negative and significant for scores in Task 1, and positive and significant for scores in Task 2, but not

	Task 1		Task 2	
	Scores	Percentiles	Scores	Percentiles
	(1)	(2)	(3)	(4)
S1 score	-0.03		-0.36	
	(0.09)		(0.33)	
S1 percentile		0.06		-0.19
		(0.15)		(0.12)
ΔS^{S2S1}	0.55***		0.23	
	(0.16)		(0.28)	
ΔP^{S2S1}		1.21***		0.95***
		(0.25)		(0.11)
$HARD$	-0.05	-0.097	0.81	0.05
	(0.52)	(0.063)	(0.74)	(0.04)
FB	3.63***	0.31*	-6.71	0.03
	(1.29)	(0.16)	(6.39)	(0.10)
(S1 score)$\cdot FB$	-0.47***		0.53	
	(0.17)		(0.43)	
(S1 percentile)$\cdot FB$		-0.54**		0.04
		(0.25)		(0.16)
$\Delta S^{S2S1} \cdot FB$	-0.39*		0.83**	
	(0.22)		(0.39)	
$\Delta P^{S2S1} \cdot FB$		-0.53		0.07
		(0.33)		(0.15)
Intercept	-1.14	-0.003	4.58	0.13*
	(0.84)	(0.103)	(4.77)	(0.07)
N	67	67	72	71
R^2	0.39	0.52	0.31	0.84

Table 9: Estimation results for the change in the overestimation of scores and percentiles between Stage 1 and Stage 2 (standard errors in parentheses). The results are obtained by estimating Eq. (3) and similar equations for each column by OLS. Significance levels: *** - $p < 0.01$, ** - $p < 0.05$, * - $p < 0.1$.

for percentiles in both tasks. For scores, this result implies that in Task 1 the students who improved their performance reacted less to feedback than those who did not, whereas in Task 2 the students who improved their performance reacted more to feedback.

3.4 Discussion

The goal of Study 2 was to isolate the impact of direct calibration feedback on the unskilled-and-unaware problem. Unlike Study 1, Study 2 is a laboratory experiment with controlled stimuli materials and feedback, and thus provides a complementary research method. In Study 2, subjects performed two tasks in two stages. Task 1 was a mathematical skill task, while Task 2 was a general knowledge task. Stage 1 was conducted at the beginning of the semester, concurrently with prediction M1 of Study 1. Stage 2 was conducted during the last week of the semester, around the same time as prediction F of Study 1. At the beginning of Stage 2, half of the subjects received direct calibration feedback about their absolute and relative performance in Stage 1.

In Task 1, we found that direct calibration feedback improved calibration, and the unskilled who received feedback improved their calibration more. In Task 2, however, we did not find any effect of feedback on calibration. In the remainder of this section, we discuss our findings for each task.

3.4.1 Task 1

The initial pattern of miscalibration in both scores and percentiles in Task 1 is consistent with the unskilled-and-unaware problem. Unlike in Study 1, subjects are relatively well-calibrated in scores already at Stage 1. We attribute this to the representativeness and familiarity of stimuli, and to the fact that calibration was measured through postdictions. These factors appear to play no role for percentiles, however, where initial miscalibration is roughly of the same magnitude as in Study 1.

The evolution of miscalibration between stages in Task 1 is different from that observed in Study 1. In Task 1, individual performance mixing between stages is the major contributor to changes in calibration. Of main interest, however, is the effect of the controlled calibration feedback half of the subjects received. We find that calibration feedback leads to an overall reduction in overconfidence. Additionally, the unskilled who received feedback reduced overconfidence more.

By construction, there was no relevant environmental feedback between stages in Task 1. Thus, the fact that we did not find the effect of Stage 1 performance on the improvement in calibration between stages for those subjects who did not receive calibration feedback

confirms the importance of environmental feedback for the results of Study 1.

Adding numbers is a very simple and straightforward task. We observed a large variation in skill, but this variation was mainly due to differences in speed, and not so much due to differences in accuracy (correlation between the number of attempted answers and the number of correct answers is 93%, indicating that there is little variation in the accuracy rate). In these circumstances, the number of attempted answers is a good predictor of the number of correct answers, which explains the good calibration in absolute performance postdictions. In terms of speed, however, this is not an easy task. More than 80% of the subjects did not solve even half of the proposed problems, and nobody solved more than 90%. In this respect, performance in Task 1 is somewhat similar to the exams in Study 1. To the extent that the results for Task 1 can be useful in explaining the results of Study 1, our findings indicate that direct calibration feedback improves calibration overall and helps mitigate the unskilled-and-unaware problem.

3.4.2 Task 2

Task 2 turned out to be exceptionally easy for our subjects, especially at Stage 1 where more than half of the subjects answered more than 80% of questions correctly, and about 20% of the subjects answered more than 90% of questions correctly. This explains the observed underconfidence of the skilled in scores. For percentiles, however, the patterns of miscalibration at both stages in Task 2 are very similar to those of Task 1.

For scores, initially, we observe a reversal of the unskilled-and-unaware problem: the unskilled are almost perfectly calibrated, while the skilled are underconfident. For percentiles, however, we do not observe any reversal. Given that Stage 1 of Task 2 was easy, this result is at odds with Burson et al. (2006), and confirms the finding of Ehrlinger et al. (2008) regarding the prevalence of the unskilled-and-unaware problem for all levels of task difficulty.

Similar to Task 1, we observe a substantial mixing of performance between the two stages. However, there is no significant effect of direct calibration feedback on the change in overconfidence, and no differential effect of feedback for different skill levels.

Task 2 is a general knowledge task consisting of dichotomous geography questions. There is much less variation in performance than in Task 1, and the variation is mainly due to differences in accuracy (more than 96% and more than 80% of subjects attempted to answer all questions at Stage 1 and Stage 2, respectively). Most of the questions were easy, which explains the relatively good calibration in scores. For relative performance judgments, however, due to the low variation in performance, the inference problem subjects faced was more difficult than in Task 1 or Study 1. In combination with substantial

performance mixing between the stages, the environment did not facilitate responsiveness to feedback.

4 Conclusions

The goal of this paper was to explore the impact of feedback on the unskilled-and-unaware problem. Documented extensively in the previous research, the unskilled-and-unaware problem consists in the inability (and/or unwillingness) of the unskilled to internalize their incompetence. Previous research has shown that the unskilled are surprisingly stubborn in their self-assessment biases, i.e. the unskilled-and-unaware problem is robust to feedback (Hacker et al. 2000, Ferraro 2006, Ehrlinger et al. 2008). The present paper aims to understand this phenomenon better. Whether or not the unskilled can improve calibration through experience and feedback, thereby reducing the unskilled-and-unaware problem, is also an important practical question from the perspective of the optimization of education and training procedures (Smith and Dumont 1997, Stone and Opel 2000).

We have reported the results of two studies through which we examined the impact of feedback on miscalibration in various tasks and feedback conditions. In both studies, we operationalized the unskilled-and-unaware problem as a negative association between miscalibration and skill level, and measured its evolution over time. We pointed out and addressed an important methodological aspect of this measurement: subjects' competence levels and performance may change across judgments, and the effect of these changes on miscalibration should be disentangled from the effect of feedback.

In Study 1, students made judgments about their absolute and relative performance on two exams in a natural course setting. We found a strong positive effect of feedback on calibration, especially for the unskilled. Thus, the unskilled-and-unaware problem was significantly reduced. Our results provide additional insights into the roles of different types of feedback, different types of judgment, and task difficulty in this phenomenon.

During the first half of Study 1, students received environmental feedback on the subject matter, and indirect calibration feedback, but did not have access to direct calibration feedback. However, we found a significant improvement in calibration already in the first half of the study, which suggests that environmental feedback and indirect calibration feedback alone can mitigate the unskilled-and-unaware problem. In the second half of the study, all types of feedback were present simultaneously, and it is impossible to assess their impacts separately. Nevertheless, as a result of this feedback, calibration improved even more.

By comparing calibration in absolute and relative performance judgments, we found

that by the end of Study 1 the unskilled-and-unaware problem in absolute judgments was completely gone, whereas residual overconfidence of the unskilled was still present in relative performance judgments. We attribute this result to the negative impact of low placement on self-image (Koeszegi 2006, Falk et al. 2006). While low absolute performance judgments do not directly prime poor relative standing as there is always hope that others in the group performed poorly too, low relative performance judgments do. This finding suggests that the framing of calibration questions and feedback may play a significant role in the responsiveness of the unskilled to feedback. For example, if the assessment of relative standings is made indirectly through an incentivized choice decision (e.g., market entry, as in Camerer and Lovallo 1999 or Bolger, Pulford, and Colman 2008, or choice between betting on own performance or a lottery, as in Hoelzl and Rustichini 2005), it may be possible to reduce the residual overconfidence in relative standings even further.

Incidental variation in the effective difficulty of the exams between the two cohorts of students allowed us to explore the impact of task difficulty on the evolution of miscalibration. During the first half of the study, when feedback was predominantly environmental, higher task difficulty was associated with a stronger improvement in calibration in relative performance judgments. However, in the second half of the study, when direct calibration feedback became available, higher task difficulty was associated with a stronger improvement in calibration in absolute performance judgments. The dependence of the effect of task difficulty on the type of judgment suggests that the presence of calibration feedback interacted with task difficulty. We attribute the difference in relative performance judgments to reference group neglect (Moore and Cain 2007): with environmental feedback, students received information on the nature and difficulty of the task, but there was no new information to base their absolute performance judgments on. In contrast, in the second half of the semester, performance feedback led to the differential effect of task difficulty on absolute performance judgments, but there was no longer any difference in relative performance judgments between cohorts.

The motivation for Study 2, a laboratory study with controlled stimuli materials and feedback, was to separately explore the effect of direct calibration feedback on the unskilled-and-unaware problem. The tasks in Study 2 – a number addition task and a geography general knowledge task – were chosen so as to minimize the effect of other types of feedback. In Task 1, we found a strong positive effect of feedback on reduction in overconfidence. We also found that the unskilled benefited more from the feedback. Combined with the results of Study 1, these findings suggest that both environmental and calibration feedback can mitigate the unskilled-and-unaware problem, and the unskilled

are not doomed to be especially unaware. In Task 2, due to its extreme simplicity and low variation in performance at Stage 1, we did not identify any effect of feedback.

The central result of our study is the fact that the unskilled-and-unaware problem can be reduced through naturally occurring interactions and feedback in a real-world environment. This finding is in contrast to the results of prior studies that reported that the unskilled are unaware not only initially but also after feedback. One of the explanations we propose is that prior studies did not fully control for the impact of changes in performance on miscalibration. It is unlikely, however, that not taking this effect into account would reverse our results, as demonstrated in Figure 1: although there is some difference in miscalibration patterns between m-cohorts and f-cohorts in prediction F, both show a significant reduction in the unskilled-and-unaware problem.

Another explanation can be the highly selective nature of our sample. The preparatory semester students at CERGE-EI are not representative students by any measure; most of them are likely to have been at the top of the skill distributions in classes at their home universities. If, as proposed by Kruger and Dunning (1999), calibration is related to metacognitive ability, it is not surprising that these students improve calibration even without any direct performance feedback. At the same time, the "unskilled" students in our study are also the ones to experience the most severe blow to their self-image. The interaction of these two factors could produce the observed difference in residual miscalibration between absolute and relative performance judgments.

Acknowledgements

We are grateful to the editor and three anonymous referees for their extensive and insightful comments, which helped improve the paper significantly.

References

Alicke M.D., & Govorun O. (2005). The better-than-average effect. In M.D. Alicke, D.A. Dunning, & J.I. Krueger (Eds.), *The self in social judgment. Studies in self and identity* (pp. 85–106). New York, NY: Psychology Press.

Arkes H.R., Christensen C., Lai C., & Blumer C., (1987). Two methods of reducing overconfidence. Organizational Behavior and Human Decision Processes, 39, 133-144.

Bolger F., Pulford B.D., & Colman A.M., (2008). Market entry decisions: Effects of absolute and relative confidence. Experimental Psychology, 55, 113-120.

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

Camerer C., & Lovallo D. (1999). Overconfidence and excess entry: An experimental approach. American Economic Review, 89, 306-318.

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

Chen G., Kim K.A., Nofsinger J.R., & Oliver M.R. (2007). Trading performance, disposition effect, overconfidence, representativeness bias, and experience of emerging market investors. Journal of Behavioral Decision Making, 20, 425-451.

Dhami K.M., Hertwig R., & Hoffrage U., (2004). The role of representative design in an ecological approach to cognition. Psychological Bulletin, 130, 959-988.

Dunning, D. (2005). *Self-insight: Roadblocks and detours on the path to knowing thyself.* New York: Psychology Press.

Dunning, D., Heath, D., & Suls, J. M., (2004). Flawed self-assessment: Implications for health, education, and the workplace. Psychological Science in the Public Interest, 5, 69-106.

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, 98-121.

Ferraro J.P., (2006). Know thyself Incompetence and overconfidence. Experimental Laboratory Working Paper Series no. 2003-001, Dept. of Economics, Andrew Young School of Policy Studies, Georgia State University.

Falk, A., Huffman, D., & Sunde, U., (2006). Self-confidence and search. IZA (Institute for the Study of Labor) Discussion Paper No. 2525.

Gigerenzer G., Hoffrage U., & Kleinboelting H., (1991). Probabilistic mental models: A Brunswikian theory of confidence. Psychological Review, 98, 506-528.

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. Journal of Educational Psychology, 92, 160-170.

Haun, D. E., Zeringue, A., Leach, A., & Foley, A. (2000). Assessing the competence of specimen-processing personnel. Laboratory Medicine, 31, 633-637.

Hoffrage, U., (2004). Overconfidence, in Pohl, R. (Ed.), *Cognitive Illusions: a handbook on fallacies and biases in thinking, judgment and memory.* Psychology Press.

Hoelzl, E., & Rustichini, A., (2005). Overconfident: do you put your money on it? The Economic Journal, 115, 305-318.

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.

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

Koriat A., Lichtenstein S., & Fischhoff B., (1980). Reasons for confidence. Journal of Experimental Psychology: Human Learning and Memory, 6, 107118.

Krajč M., & Ortmann A., (2008). Are the unskilled really that unaware? An alternative explanation. Journal of Economic Psychology, 29, 724-738.

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.

Lee, J.-W., Yates, J.F., Shinotsuka, H., Singh, R., Onglatco, M.L.U., Yen, N.S., Gupta, M., & Bhatnagar, D. (1995). Cross-national differences in overconfidence. Asian Journal of Psychology, 1, 63-69.

Lichtenstein S., & Fischhoff B., (1977). Do those who know more also know more about how much they know? Organizational Behavior and Human Performance, 20, 159-183.

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

Malmendier, U., & Tate, G. (2005). CEO overconfidence and corporate investment. Journal of Finance, 60, 2661-2700.

McKenzie C.R.M., (1997). Underweighting alternatives and overconfidence. Organizational Behavior and Human Decision Processes, 71, 141-160.

Moore D.A., & Cain D.M., (2007). Overconfidence and underconfidence: When and why people underestimate (and overestimate) the competition. Organizational Behavior and Human Decision Processes, 103, 197-213.

Moore D.A., & Healy P.J., (2008). The trouble with overconfidence. Psychological Review, 115, 502-517.

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, 1067-1101.

Pulford B.D., & Colman A.M., (1997). Overconfidence: feedback and item difficulty effects. Personality and Individual Differences, 23, 125-133.

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

Sharp G.L., Cutler B.L., & Penrod S.D., (1988). Performance feedback improves the resolution of confidence judgments. Organizational Behavior and Human Decision Processes, 42, 271-283.

Sieck W.R., Arkes H.R., (2005). The recalcitrance of overconfidence and its contribution to decision aid neglect. Journal of Behavioral Decision Making, 18, 29-53.

Sieck W.R., Merkle E.C., & Van Zandt T., (2007). Option fixation: A cognitive contributor to overconfidence. Organizational Behavior and Human Decision Processes, 103, 68-83.

Smith, D., & Dumont, F., (1997). Eliminating overconfidence in psychodiagnosis: strategies for training and practice. Clinical Psychology: Science and Practice, 4, 335-345.

Sniezek, J. A., & Buckley, T. (1991). Confidence depends on level of aggregation. Journal of Behavioral Decision Making, 4, 263-272.

Stone E.R., & Opel R.B., (2000). Training to improve calibration and discrimination: The effects of performance and environmental feedback. Organizational Behavior and Human Decision Processes, 83, 282-309.

Whitcomb, K. M., Önkal, D., Curley, S. P., & Benson, P. G. (1995). Probability judgment accuracy for general knowledge: Cross-national differences and assessment methods. Journal of Behavioral Decision Making, 8, 51-67.

Wooldridge J., (2002). *Econometric analysis of cross section and panel data.* MIT press.

Wright, G. N., Phillips, L. D., Whalley, P. C., Choo, G. T., Ng, K. O., Tan, I., & Wisudha, A. (1978). Cultural differences in probabilistic thinking. Journal of Cross-Cultural Psychology, 9, 285-299.

Yates, F., Lee, J-W., & Bush, J.G., (1997). General knowledge overconfidence: Cross-national variations, response style, and "Reality." Organizational Behavior and Human Decision Processes, 70, 87-94.