ARE THE UNSKILLED REALLY THAT UNAWARE? UNDERSTANDING SEEMINGLY BIASED SELF-ASSESSMENTS

Marian Krajč

CERGE-EI

Charles University
Center for Economic Research and Graduate Education
Academy of Sciences of the Czech Republic
Economics Institute

WORKING PAPER SERIES (ISSN 1211-3298)
Electronic Version

373

Working Paper Series 373
(ISSN 1211-3298)

Are the Unskilled Really That Unaware? Understanding Seemingly Biased Self-Assessments

Marian Krajč

CERGE-EI
Prague, November 2008

ISBN 978-80-7343-173-0 (Univerzita Karlova. Centrum pro ekonomický výzkum a doktorské studium)
ISBN 978-80-7344-162-3 (Národohospodářský ústav AV ČR, v.v.i.)

Are the Unskilled Really That Unaware? Understanding Seemingly Biased Self-Assessments

Marian Krajč

CERGE-EI[*]

Abstract

The so-called unskilled-and-unaware problem was experimentally identified a decade ago: The unskilled are seemingly afflicted by a double curse because they also seem unaware of their (relative) lack of skills. Numerous authors have elaborated on this problem – experimentally as well as theoretically. In this paper, we report on the results of three experiments (one field, two laboratory) through which we test a theoretical model and some informal extensions. Specifically, we examine the impact of general information and specific information (feedback) on the quality of self-assessment ("calibration") in various tasks and under various conditions. Overconfidence behavior initially prevails in almost all settings. We find a strong positive effect of general information on calibration, and show that calibration improves more when feedback is provided. In our experiments, it is the unskilled who improve their calibration the most.

Abstrakt

Takzvaný *unskilled-and-unaware* problém bol experimentálne identifikovaný pred desaťročím: *Unskilled* sú zdanlivo postihnutý dvojitým prekliatím pretože sa zdajú byť nevedomí si nedostatku svojich schopností (v porovnaní s inými). Viacerí autori rozvinuli tento problém – experimentálne aj teoreticky. V tomto článku referujeme výsledky troch experimentov (jedného z terénu a dvoch laboratórnych) pomocou ktorých testujeme teoretický model a niekoľko neformálnych rozšírení. Konkrétne, skúmame vplyv všeobecnej informácie a špecifickej informácie (spätnej väzby) na kvalitu sebaohodnotenia („kalibrácie") v rôznych úlohách a za rôznych podmienok. Prílišná sebaistota spočiatku prevláda v takmer všetkých situáciách. Identifikujeme silný pozitívny efekt všeobecnej informácie na kalibráciu a ukážeme, že kalibrácia sa zlepší ešte viac keď je poskytovaná spätná väzba. V našich experimentoch sú to práve *unskilled* ktorí si najviac zlepšia svoju kalibráciu.

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

[*] CERGE-EI, a joint workplace of the Center for Economic Research and Graduate Education at Charles University and the Economics Institute of the Czech Academy of Sciences.

Acknowledgements: I am grateful to Ralph Hertwig, Peter Katuščák, Petr Zemčík, and participants of the 8th GDN Regional Research Conference for their comments. Special thanks to Andreas Ortmann and Dmitry Ryvkin for advice and guidance. I would also like to thank Aldo Rustichini for his help during my mobility stay at the University of Minnesota.

This research was supported by a grant from the CERGE-EI Foundation under a program of the Global Development Network. All opinions expressed are those of the author and have not been endorsed by CERGE-EI or the GDN.

1. Introduction

The unskilled-and-unaware problem was first identified by psychologists Kruger and Dunning (1999). These authors conducted several experiments, mostly with students, in which they identified the following three regularities: people ranked at the bottom of the skills distribution overestimate their relative ranking, those ranked at the top of the skills distribution underestimate their relative ranking, and these alleged miscalibrations are asymmetric – many more unskilled underestimate their relative standing and often do so quite dramatically. These three observations apply to relative rankings as well as absolute score measures. Kruger and Dunning (1999) focus on the case of relative rankings and draw the conclusion that the unskilled lack the metacognitive ability to realize their incompetence. A number of studies were subsequently written on the unskilled-and-unaware problem, both experimental and theoretical.

In the current paper, we experimentally test the assumptions and the performance of a theoretical model (Krajč and Ortmann, 2008) that explains the unskilled-and-unaware problem with asymmetric distribution of skills and errors in judgment. We also investigate the impact of various types of information on calibration (and on the magnitude of the unskilled-and-unaware problem). In addition, we compare calibration in relative and absolute self-assessment. Lastly, we try to answer the question proposed by Juslin et al (2000): What is the relationship between calibration in general knowledge-oriented tasks and calibration in skill-oriented tasks?

The results of our three experiments (one field experiment and two embedded laboratory experiments) suggest, on average, mostly overconfident behavior. The results provide some support for the assumptions of the theoretical model proposed by Krajč and Ortmann (2008) but the model's performance is, for various reasons, not as impressive as expected. We also show that information improves calibration, especially of the unskilled, and hence reduces the unskilled-and-unaware problem. Moreover, we identify weakly better calibration in skill-oriented than in general knowledge-oriented tasks, thereby shedding some light on the question identified by Juslin et al (2000) as being in need of an answer.

The present paper is organized as follows. In Section 2, we review the literature concerned with the unskilled-and-unaware problem and related issues. In Section 3, we motivate, detail, and enumerate our research objectives and research strategy. Section 4 describes the design and implementation of the experiments. In Section 5, we present the results. We discuss our results and conclude in Section 6.

2. Literature review

The results and conclusions of Kruger and Dunning (1999) prompted a flurry of critical studies. For example, Krueger and Mueller (2002) showed that the use of unreliable measures[1] of ability can lead to results similar to those reported in Kruger and Dunning (1999). Specifically, the authors showed that the unreliability of measures essentially causes the measured ability to regress toward the mean (also known as regression-to-the-mean), which induces overestimation (underestimation) in the lower (upper) part of the distribution. In addition, to explain the asymmetry, the authors used the presence of the better-than-average effect. However, we submit that the better-than-average effect used by Krueger and Mueller (2002) to explain the asymmetry is itself the result of people's behavior and should not be used as an explanatory element.

Burson et al (2006) were concerned with the asymmetry and tried to explain it by introducing task difficulty into the unskilled-and-unaware problem. The authors experimentally showed that the degree of over- and underestimation depends on the task difficulty. Indeed, their results were similar to those of Kruger and Dunning (1999) for easier tasks (with asymmetry in over- and underestimation) yet they showed less overestimation of unskilled and more underestimation of skilled for harder tasks. Actually, asymmetry in over- and underestimation disappeared (or even was reversed – more underestimation among the skilled than underestimation among the unskilled) in experiments with harder tasks. Burson et al (2006) concentrated mostly on the unskilled-and-unaware problem under the percentile estimation and also made an effort to control for unreliability of percentile estimation.

[1] Percentile is not perfectly reliable measure of abilities. Lack of reliability in a test makes the highest performers look less able than they are and the poorest performers less deficient than they are.

Ehrlinger et al (2008) addressed objections and suggestions (to the results and experimental setup in Kruger and Dunning, 1999) of various critical studies. Mainly, the authors used financial and social incentives and real-world situations, and also controlled for unreliability of measures. In spite of these changes, the pattern observed in Kruger and Dunning (1999) survived (overestimation of their skills by the unskilled and underestimation of their skills by the skilled, and miscalibration much more dramatic for the unskilled than the skilled). Moreover, Ehrlinger et al (2008) also tried to identify the cause of this pattern of miscalibration in percentile ranking. The improvement in calibration was found to be stronger when the authors corrected for the errors in people's own raw score than when they corrected for misperception about others; the improvement in calibration of the skilled was found to be approximately the same when they corrected for the errors in people's own raw score and when they corrected for misperception about others.

Recently, Krajč and Ortmann (2008) offered an alternative explanation of the unskilled-and-unaware problem: They constructed a simple model that shows that the unskilled, rather than being more unaware than the skilled, face a tougher inference problem which, at least partially, explains the alleged lack of metacognitive ability. The Krajč and Ortmann (2008) model is based on two assumptions. First, they claim that the distribution of students' skills[2] is bounded and that skills have J-distribution[3]. Second, the authors assume that the self-assessment process involves unsystematic noise[4]. Employing these two assumptions, the authors generated through computational simulations patterns of miscalibration similar to those reported by Kruger, Dunning, and their collaborators, and showed that people do not have to be necessarily miscalibrated to produce behavior consistent with these patterns; the unskilled may simply have a tougher inference problem than the skilled. The authors also showed that the first

[2] Cornell and Chicago university students are typically used in studies of the unskilled-and-unaware problem.

[3] By J-distribution the authors mean a distribution with greater mass in the left (the unskilled) than in the right (the skilled) tail of the distribution. The authors justified why the samples used in earlier studies should satisfy this assumption.

[4] This assumption is often used in the literature (e.g., Erev et al, 1994). It can be justified as follows: "Noise is likely to be correlated with familiarity (and hence feedback about one's own standing) with a particular domain. If one is not that familiar, one is likely to use one's self-assessment from other domains as a proxy, which adds to the error." (Krajč and Ortmann, 2008, pp. 729).

assumption can to some extent be weakened, while the qualitative results remain the same

Krajč and Ortmann (2008) also discussed the conditions under which they expect the unskilled-and-unaware problem to disappear. The authors pointed out the importance of the distribution of real abilities, task randomness, diagnosticity of feedback, and use of real financial and other incentives in the research on (the lack of) metacognitive ability. It is well understood by most experimental economists that every experimental test is always a joint test of the theory that is being tested and the specific way the experiment is implemented (Duhem-Quine hypothesis; see Smith, 2002).

3. Motivation and research objectives

The unskilled-and-unaware problem is likely to show up in all those real-world situations where self-assessment (relative ranking or absolute assessment) matters. For example, biased self-assessments could cause managers to undertake inappropriate projects, or biased self-assessments could create problems on the labor market for workers and the unemployed (like extension of the waiting time for a job with negative consequences on the most vulnerable group). In addition, the alleged unawareness of the unskilled could lead, through excessive expectations, to disappointment and frustration and thus have a negative psychological impact.

If the unskilled-and-unaware problem applied also to market entry (games), an excessive entry of the unskilled and an insufficient entry of the skilled would be observed. Camerer and Lovallo (1999) introduced the skill-dependent ranking and rank-dependent payoff in a market entry experiment. These authors showed that people excessively enter the market when their payoff depends on their relative skills and concluded that they overestimate their abilities. This finding could have an important impact on entrepreneurship and market entry. In trying to understand whether the Duhem-Quine critique applies to Camerer and Lovallo (1999), Elston et al (2006) showed that neither non-entrepreneurs nor entrepreneurs are overconfident about their skills in market entry games. In contrast, wannabe-entrepreneurs are. There are at least three possible explanations for the contradictory results of Camerer and Lovallo (1999) and Elston et al (2006). First, subjects might have had different distribution of abilities in these experiments. Second, the overconfidence bias might be specific for some narrow group of people. Third, Camerer and Lovallo (1999) may not have had enough relevant control variables such as measures of risk aversion and desire to win, which Elston et al (2006) used. The contradictory results of these two sets of authors suggest, in any case, the importance of the choice of the subject pool and distribution of skills in this pool on miscalibration. Knowing the true source of the identified miscalibration could help the afflicted to avoid it. A similar logic also applies to various other areas (managers undertaking projects, workers asking for promotion, unemployed searching for a new job, etc.).

The findings of Krajč and Ortmann (2008) suggest that currently available results on the unskilled-and-unaware problem, as well as those involving self-assessments more generally, are likely to lead to misleading conclusions and policy recommendations if they do not deal with the issue of the subject pools (and most of them do not). Miscalibration in self-assessment may be caused by something other than non-rational behavior. Since the existing literature does not deal with this issue, with our experiments we try to shed more light on it. Concretely, our goal is to test the theoretical model proposed by Krajč and Ortmann (2008). We test the assumptions of the model (J-distributed score and error in judgment) as well as its performance in generating ability perception.

Hypothesis 1: *The* ***model*** *in Krajč and Ortmann (2008) generates patterns of miscalibration similar to the predictions/estimates of one's own score observed in the experiment.*

The main aim of our experiments, however, is to identify the impact of information on calibration in absolute and relative self-assessment. Various authors have shown that better information can lead to better judgments and decisions (e.g., Duffy and Hopkins, 2005, Engelmann and Strobel, 2000). Krajč and Ortmann (2008) therefore conjecture that information about one's own ability and abilities (and their distribution) of others plays an important role, especially in relative self-assessment. Our working hypothesis is that a substantial part of the miscalibration typically reported stems from insufficient information about the subject pool or the task.[5] Specifically, to explain the impact of various types of information on miscalibration (unskilled-and-unaware problem) we test two hypotheses.

Hypothesis 2a: ***General information*** *decreases miscalibration.*

Hypothesis 2b: *Lower miscalibration with specific information (**feedback**) than without it.*

In our experiments we also try to identify the relationship between the absolute and relative self-assessment in various situation and types of tasks. Notwithstanding numerous studies on absolute or relative self-assessment (e.g., Kruger and Dunning, 1999, Burson et al, 2006, Juslin et al, 2000 – for a review), to the best of our knowledge no one seems to have investigated the relationship between absolute and relative self-assessment and a possible causality before. It is possible that people first estimate their own absolute score and the group quality and then infer their relative position. On the other hand, it is also possible that people do not take estimates of their own score into account and create their percentile estimates based on something else. Due to the fact that people, when evaluating their relative ranking, have to assess their own absolute ability as well as the ability of others (or at least the number of people with better/worse ability), relative self-assessment seems to be a more complicated problem. If true, we should observe lower calibration in the relative than in the absolute self-assessment. Comparing miscalibration in absolute and relative measures is the first step in answering the question how people create their estimates (absolute and relative).

Hypothesis 3: *There is less miscalibration in own score estimates than in percentile estimates (**Own score vs. Percentile**).*

Moreover, we investigate the relationship of over/underconfidence in self-assessment tasks and general-knowledge tasks. We conjecture that the ability perception in skill-oriented tasks might differ from the perception in knowledge-oriented tasks. Juslin et al

[5] We did a pilot experiment with prep students in 2004, in which we identified the unskilled-and-unaware problem in that data. The magnitude of the effect decreased with more information available to students (toward the end of the semester).

(2000)[6] have identified this question as one in need of an (empirical) answer. Earlier, Klayman et al (1999) showed that the degree of overconfidence varied over domains, yet was not a function of domain difficulty. Therefore, comparing two major domains, as the general knowledge-oriented and skill-oriented tasks are, could initiate a discussion about (possible) differences in these two domains. There is a large body of literature on general-knowledge questions and on skill self-assessment tasks, yet a direct comparison of calibration in these two types is lacking.

Hypothesis 4: *Skill-oriented tasks generate less miscalibration than general knowledge-oriented tasks (***skills vs. general knowledge***).*

The issue of representativeness of stimuli in experiments also plays an important role in psychological studies on overconfidence (e.g., Gigerenzer et al, 1991, Juslin et al, 2000, Dhami et al, 2004). For example, Juslin et al (2000) showed, with a metastudy, that the so-called hard-easy effect (overconfidence more common for hard and underconfidence for easy item samples) in general-knowledge questions typically appears only in studies that use non-representative sampling (e.g., selected alternatives). So does overconfidence. In our experiments, we try to control for this in the laboratory experiments.[7]

4. Experiments

To test the hypotheses, we conducted three experiments (Experiments 1, 2, and 3). All three experiments were designed to address partially overlapping subsets of our hypotheses. The first experiment (Experiment 1) was a field experiment of sorts: we compared midterm and final exam predictions of a newly constituted class in a real-world setting. The second and third experiments, laboratory experiments of sorts, were embedded in the field experiment and addressed all four research hypotheses.

[6] Juslin et al (2000), p. 394 express a need to make progress towards answering the open question of the relation between overconfidence as exhibited in self-assessment tasks and general-knowledge tasks.

[7] Originally, we actually formulated this as the hypothesis “Less miscalibration in laboratory experiments (with representative stimuli) than in field experiment (without representative stimuli).” We decided not to test this hypothesis because there are other factors that might contribute to this difference. For example, the estimates in laboratory experiments were made after the task while in field experiment before the task. Moreover, overconfidence varies across domains and it therefore can be different in our lab experiments than in field experiment.

Table 1. *Hypotheses tested in Experiments 1, 2, and 3.*

	H1	H2a	H2b	H3	H4
Experiment 1	✓	✓	×	✓	×
Experiment 2	✓	✓	✓	✓	✓
Experiment 3	✓	✓	✓	✓	✓

Note that none of the three experiments were marred by subject selection problems as our participants were "pseudo-volunteers" (see Eckel and Grossman, 2000). In other words, the selection process that brings them to the experiments is unrelated to the experimental tasks. A possible disadvantage with pseudo-volunteers is that the subjects may simply not be interested in participating in the experiment (Harrison and Rutstroem, 2007, especially fn 79). Given the time our experiments took and the substantial financial incentives we provided, as well as our observation of our pseudo-volunteers' conduct, we do not believe that we have to worry much about this possible disadvantage.

4.1. Experiment 1

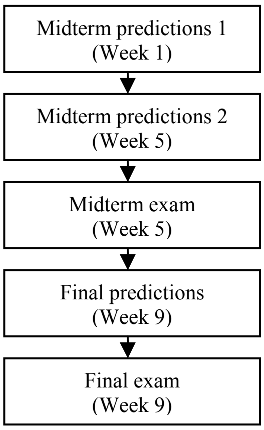
With this experiment we addressed research hypotheses 1, 2a, and 3.

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

4.1.1. Design

In Experiment 1, we asked students of the micro course in the prep semester at CERGE-EI to predict their performance in the micro course[8] and the average score as well as the percentage of better performing students on both the midterm and final exam. Students made these predictions twice for the midterm exam (in the very first week of prep before Stage 1 of Experiment 2 and right before the midterm) and once for the final exam (right before the final).[9] Figure 1 illustrates the basic structure of Experiment 1.

Figure 1. *The structure of Experiment 1.*



4.1.2. Implementation

A total of 49 (52) students of the prep semester at CERGE-EI made their predictions about midterm performance in the very first week of the prep semester (right before the midterm) and 45 (51) of these students participated in the midterm exam. Altogether 53 students sat for the midterm exam.[10] A total of 46 students made their predictions about final performance right before the final and 45 of them took the final exam. Altogether 46 students sat for the final exam.[11] For each question, the participant with the best prediction was rewarded with 500 CZK.[12] All participants were told that their predictions would not affect their grades and that no one but the researchers would see the data.

[8] Ferraro (2005) used in-class exams to study the relationship between self-awareness and overconfidence, where students evaluated their absolute score and relative standing on three in-class multiple-choice exams. After each exam, they received feedback (score, mean, median, letter grade frequencies). The main advantage of our experiment is that our subject pool is newly formed which allows us to observe evolution of calibration as students get know each other.

[9] Complete instructions to Experiment 1 are available on request from the author.

[10] 2 students came to the midterm exam after the questionnaire with predictions had been collected.

[11] 1 student did not hand in the exam sheet and 1 student came late.

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

We asked our subjects to predict their performance[13] ("What is your prediction of your own score on the midterm exam in microeconomics?"), average score ("What is your prediction of the average score on the midterm exam across all those who take the microeconomics exam in prep semester?") as well as percentile ranking[14] ("What do you think is the percentage of people in the group who will perform better than you on the midterm exam in microeconomics?") on the micro midterm and final exam. The first question allowed us to measure estimated score (perceived ability) and thus over- and underestimation of subjects' own score (ability). We also were able to compute, by means of the theoretical model in Krajč and Ortmann (2008), the perceived ability from the real score distribution (real ability) and compare it to perceived score ability distribution from the experiment in order to see whether the model generates similar patterns as the experiment (hypothesis 1: "*model*").

The second question revealed some information about participants' beliefs about the quality of the group combined with the task difficulty. With the third question we measured percentile ranking as Kruger and Dunning (1999) did. As Experiment 1 was conducted at three different points in time, it allowed us to observe the evolution of the level of miscalibration in Own score, Average score, and Percentile over time (hypothesis 2a about the effects of "*general information*").

We were also able to compare calibration in Own score and Percentile predictions (hypothesis 3: "*Own score vs. Percentile*") and analyze how this relationship evolves over time.

Note that Experiment 1 is interesting not only due to the real-world setting with high stakes (for prospective PhD students) but also due to the feedback type – natural feedback for the given situation. This means that we did not give our subjects any artificial feedback; they only received natural feedback from the course (like homework grades, midterm results, midterm distribution) that was directly connected to the task as well as indirect feedback from other classes (macro, mathematics). This is a real-world

[13] We will call this measure "Own score" throughout the paper.
[14] We will call this measure "Percentile" throughout the paper.

situation where feedback and self-evaluation matter. We can find many other real-world situations like this (e.g., all types of students at their schools, employees working in a team, participants in various retraining courses).

4.2. Experiments 2 and 3

With these experiments we addressed hypotheses 1 through 4.

4.2.1. Design

Experiment 2 and Experiment 3 were laboratory experiments that were conducted together (one after another). Each experiment was conducted in two stages (Stage 1 and Stage 2) and in each experiment we used a different task.

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

Task, Experiment 3: Participants had to complete, within a 2-minute time limit, a quiz containing 20 two-alternative general-knowledge questions, a research strategy widely investigated in psychology.[15] In Stage 1, we asked for a comparison of the population of pairs of European Union countries ("Which of the following two countries has a larger population?") while in Stage 2, in order to avoid learning effect, we asked for a comparison of the population of pairs of the 50 most populated world countries.[16] This task is a general knowledge-oriented task (knowledge of geography).

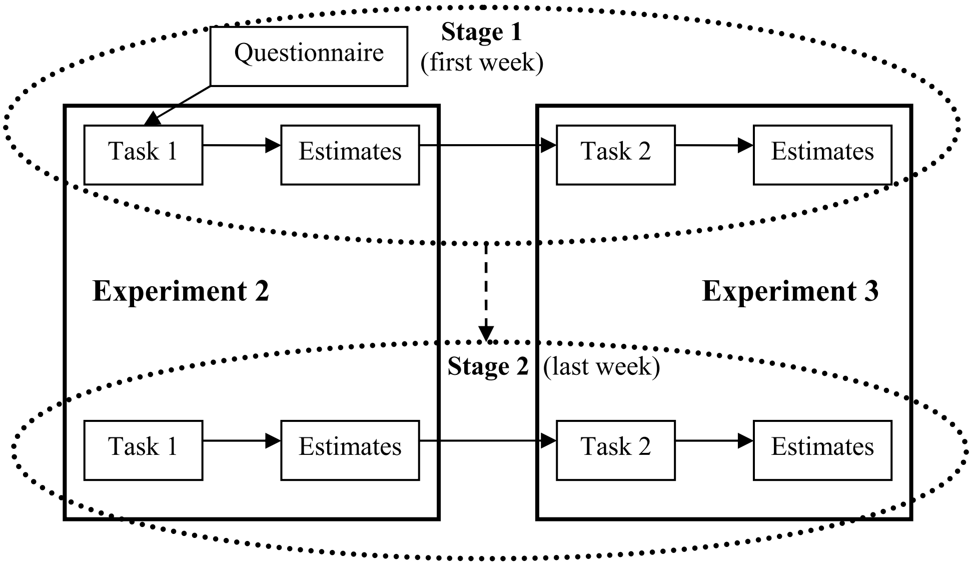
Participants were also asked to answer some self-evaluative questions (described below in more detail) after performing the corresponding task. The structure of Experiments 2 and 3 is depicted in Figure 2.[17]

[15] E.g., for a review see Juslin et al (2000).

[16] By the learning effect we mean that some people, motivated by Stage 1 of the experiment, could learn the population of these countries and thus we would artificially change the knowledge and might get non-representative data.

[17] Complete instructions to Experiments 2 and 3 are available upon request from the author.

Figure 2. *The structure of Experiment 2 and Experiment 3.*



4.2.2. Implementation

A total of 49 (45) students of the prep semester at CERGE-EI participated in Stage 1 (Stage 2) of Experiments 2 and 3. Stage 1 (Stage 2) lasted 25 (20) minutes. All participants were paid according to their performance in the experiment. The average payoff was 177 CZK (313 CZK) in Stage 1 (Stage 2).

In order to measure the effect of overall information on self-assessment, Experiments 2 and 3 consisted of two stages. Stage 1 was conducted at the very beginning (first week) of the prep semester while Stage 2 at the end (last week) of the prep semester when students could be assumed to have more information about their relative standing in the group (hypothesis 2a: "*general information*"). We did not tell our subjects that Stage 2 would follow. All instructions were read aloud.

Stage 1: After providing a brief general introduction to the experiment, we asked our subjects to fill in a short questionnaire (age, sex, and background – mathematician or economist). We then continued with instructions to Experiment 2: we explained that the task is to sum 5 two-digit numbers and gave an example. The subjects were also informed that for each correctly solved problem they would be paid 5 CZK. Afterwards, we distributed sheets with 22 summing problems and gave our subjects 3 minutes to solve as many of these problems as possible. Finally, similarly as in Experiment 1, we asked subjects to provide estimates of their score, relative percentile ranking, and group average score. The most accurate estimate to each of these questions was rewarded with 500 CZK. Then, Experiment 3 with an identical procedure, but a different task, followed.[18]

Stage 2, Experiment 2: Similar to Stage 1, but with the following changes. First, we increased the incentives to 10 CZK for each correctly solved summing problem.[19] Second, in Stage 2, one half of the participants received for each task feedback about their performance (own score, average score, and percentage of better scoring people) in Stage 1.

Stage 2, Experiment 3: Similar to Stage 1, but with the following changes. First, in order to avoid a learning effect we changed the reference class in Experiment 3: we used the 50 most populated world countries (instead of European Union countries). Second, in Experiment 3 we gave our subjects 40 general-knowledge questions, keeping the reward for a correct answer constant (5 CZK)[20]. Third, in Stage 2, one half of the participants received for each task feedback about their performance (own score, average score, and percentage of better scoring people) in Stage 1.

In both experiments, people for the feedback treatment were randomly selected (in a stratified manner)[21] just before each task. Therefore, in addition to some indirect (natural) feedback acquired from the micro, macro, and math results from the midterms and homework, some subjects received direct feedback about the performed tasks.

[18] I.e. instructions explaining that the task was to compare populations of pairs of European Union member countries, rewarding 5 CZK for each correct answer to each of 20 pairs of countries they were asked to compare within 2 minutes, and Own score, Average score, and Percentile estimates.

[19] Because time gets scarcer towards the end of the semester, we decided to increase the incentives for our subjects. We doubled the reward for correct answers in Stage 2 of Experiment 2. According to the analysis of e high and very high payoff treatments in Rydval and Ortmann (2004) this should not matter.

[20] As answering a question in Experiment 3 was less time demanding than in Experiment 2, we doubled the reward for a correct answer in Experiment 2 and doubled the number of questions in Experiment 3. According to the analysis of high and very high payoff treatments in Rydval and Ortmann (2004) this should not matter.

[21] We split subjects according to their performance in Stage 1 into four quartiles and randomly selected half of the subjects in each quartile for the feedback treatment.

Herewith we can investigate how the strength of the feedback influences calibration of people (hypothesis 2b: "*feedback*").

In parallel to Experiment 1, we tested whether the model (Krajč and Ortmann, 2008) generates similar patterns as the experiment (hypothesis 1: "*model*"). We also compared calibration in Own score and Percentile estimates (hypothesis 3: "*Own score vs. Percentile*").

Note that the task in Experiment 2 is more skill-oriented while the task in Experiment 3 is more knowledge-oriented. We were therefore able to observe how the distributions of skills and knowledge differ and how they are related to each other, if at all (hypothesis 4: "*skills vs. general knowledge*").

In Experiments 2 and 3, unlike Experiment 1 where it was beyond the control of the experimenters, we also attempted to take into account the issue of representativeness of stimuli. We therefore used tasks that made it possible to control for this. First, we clearly specified the class of questions (so-called reference classes: all two-digit numbers, all EU countries, and 50 most populated world countries, respectively). Second, we randomly chose the numbers and the pairs of countries from the reference class. Thus, we presented our subjects with a representative sample of problems as suggested by previous research.

Incentives play an important role in various types of studies (see Camerer and Hogarth, 1999, or Rydval and Ortmann 2004, or Hoelzl and Rustichini, 2005). In Experiments 2 and 3, 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 expected that monetary incentives would help us obtain a more accurate measure of participants' abilities. To motivate the subjects to give as precise answers as possible, we thought about using a linear incentive scheme.

Since there were only about 50 people in the subject pool, we had to make a choice between using two feedback treatments or two incentive treatments. The evidence in

Cesarini et al (2006) strongly suggests that, in the present context, incentives are of lesser importance than feedback. We therefore decided to use two feedback conditions, which is not ideal but was the best we could do under the conditions that we had.

5. Results

We first report and briefly discuss the basic statistics. Then we test the hypotheses one by one, separately for each experiment. The graphs depicting the data can be found in Appendix A1 (Experiment 1) and A2 (Experiments 2 and 3).

As we are primarily interested in miscalibration, we will mostly refer to miscalibration. All statistics in Tables 2a, 2b, and 2c (mean, standard deviation) are expressed in overestimation – the difference between an estimate and the real value of the variable under investigation (Own and Average score)[22] and vice versa for percentage of better performing people (Percentile)[23]. Thus, a positive number means (on average) overestimation of the particular variable.[24] Note that while Own and Average scores are measured in scores, Percentile is measured in percentage.

Table 2a. *Experiment 1.*[25] *Basic statistics (mean, st. deviation) of overestimation of Average and Own score and Percentile is shown in columns.*

Midterm prediction 1	Average	Own	Percentile
Mean	21.52	30.07	0.23
St. Dev.	10.28	23.21	0.27
Midterm prediction 2	Average	Own	Percentile
Mean	20.28	26.30	0.20
St. Dev.	15.00	20.20	0.28
Final prediction	Average	Own	Percentile
Mean	3.98	12.85	0.11
St. Dev.	10.08	15.25	0.22

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

[23] In the case of Percentile a positive number means overestimation of own relative ranking. E.g., if one's real percentile ranking is 20 and one predicted that 10% will perform better – we observe a positive number (0.1).

[24] A few people (only 18 out of over 1,000 estimates) reported some estimates in intervals instead of numbers; we replaced these estimates with the midpoint of that interval (e.g., 50-60 with 55). Similarly, some people (only 5 out of over 1,000 estimates) gave some score predictions in percentage; we transformed these predictions to their equivalents in numbers.

[25] We excluded from the analysis those students who made predictions for an exam but did not participate in the exam (4 in midterm prediction 1, 1 in midterm prediction 2, and 1 in final prediction).

Table 2b. *Experiments 2 and 3, Stage 1.*[26] *Basic statistics (mean, st. deviation) of overestimation of Average and Own score and Percentile is shown in columns.*

Experiment 2	Average	Own	Percentile
Mean	0.51	0.62	0.11
St. Dev.	3.07	1.65	0.30
Experiment 3	Average	Own	Percentile
Mean	-3.54	-2.28	0.15
St. Dev.	3.78	2.68	0.30

Table 2c. *Experiments 2 and 3, Stage 2.*[27] *Basic statistics (mean, st. deviation) of overestimation of Average and Own score and Percentile is shown in columns – all subjects, subjects with feedback, and subjects without feedback.*

Experiment 2	Average			Own			Percentile		
Feedback	pooled	feed	no feed	pooled	feed	no feed	pooled	feed	no feed
Mean	0.57	0.32	0.80	1.45	1.00	1.87	0.09	0.02	0.16
St. Dev.	3.20	3.20	3.26	2.25	1.84	2.53	0.28	0.15	0.34
Experiment 3	Average			Own			Percentile		
Feedback	pooled	feed	no feed	pooled	feed	no feed	pooled	feed	no feed
Mean	-1.36	-1.05	-1.58	0.39	-0.08	0.73	0.11	0.00	0.20
St. Dev.	3.12	3.28	3.04	3.19	2.86	3.43	0.36	0.36	0.34

Note that we transformed some data from Experiments 1 and 3[28] and we use the transformed data in all analyses below.

From Tables 2(a, b, c) we can see that, on average, overconfident behavior prevails in almost all types of predictions (except for Own and Average score in Stage 1 of Experiment 3, Average score in Stage 2 in Experiment 3, and Percentile in Final prediction). However, overconfident behavior is much stronger in the initial stage of (field) Experiment 1 than in the initial stages of the (laboratory) experiments – one strongly related to the field experiment (Experiment 2), the other one to the general-knowledge problem (Experiment 3). We observe that the mean of overestimation decreases over time (with more information) for all types of predictions made in Experiment 1. So do standard deviations. The effect of information is also evident in the treatment condition for Stage 2 of Experiments 2 and 3. We see a strong influence of information on (mis)calibration in the feedback and no feedback treatment.

[26] We excluded from the analysis one student because he/she probably misunderstood the task and computed the average, not the sum, of the given numbers.

[27] We excluded from the analysis one student because he/she probably misunderstood the task and computed the average, not the sum, of the given numbers.

[28] We transformed the Midterm predictions 1 because we asked people for estimates on a scale 0-100 but the instructor set the range of available points from 0 to 90; we multiplied the predictions by 9/10. In the case of Final predictions, we multiplied all data (predictions and score) by 3/4. In Stage 2 of Experiment 3 we asked our subjects 40 questions (unlike in other tasks, where we asked 20 questions) and therefore, to get comparable numbers, we divided all estimates and score by 2.

5.1. **Hypothesis 1** - model

The model in Krajč and Ortmann (2008) generates patterns of miscalibration similar to the predictions/estimates of Own score observed in the experiment.

5.1.1. Experiment 1

Recall that Krajč and Ortmann (2008) constructed a simple model that is built on bounded J-distribution of skills/abilities and error making in the self-assessment process. The model, essentially, imposes an idiosyncratically distributed error on all people in each real ability category (truncated for categories close to bounds) and generates a distribution of people over perceived abilities. With simulations, the authors show that this model generates patterns of miscalibration similar to those produced by preceding experiments (e.g., Kruger and Dunning, 1999) – overestimation of the unskilled and underestimation of the skilled, with an asymmetric distribution of overestimation and underestimation (the three stylized facts mentioned earlier).

The procedure in testing Hypothesis 1 in the present context is the following:

1. Create a distribution of real score (let's call it the "real distribution"): count how many people fell into each of the ability (real score) categories.[29]
2. Apply the model proposed by Krajč and Ortmann (2008)[30] on real distribution, compute the "simulated perceived score distribution" for every possible error width.
3. For each error width, compute the sum of absolute differences between the "simulated perceived score distribution" and "experimental perceived score distribution" as the discriminating measure of fit.
4. To identify the most appropriate error ("calibrated error"), select the error width with the best fit (lowest sum of absolute differences) for each prediction (Midterm predictions 1 and 2 and Final prediction).

[29] As the model hinges also on the bounds of the ability range (scores), we did the analysis only for the range of scores between the minimum of the real and estimated score and the maximum of the real and estimated score.

[30] To recall, the model assumes real distribution of score x and an error in perception ε. (Simulated) perceived distribution y is then formed as: $y = x + \varepsilon$. We did this with various widths of error.

5. Discuss the width of the calibrated error.
6. Discuss the model's ability to replicate the pattern generated by the experiment.

Table 3. *Model performance after performing steps 1 through 4.*

	# of categories*	Error width (categories)**[31]	Average deviation***
Midterm p. 1 (5)[32]	18	17	1.19
Midterm p. 2 (5)	18	17	0.97
Final p. (5)	24	23	0.70
Midterm p. 1 (10)[33]	9	9	1.18
Midterm p. 2 (10)	9	9	0.88
Final p. (10)	12	11	0.59

* - The number of ability categories over which people were distributed
** - Width of the error where the model fit is the best – calibrated error
*** - Minimum sum of absolute differences (best fit)

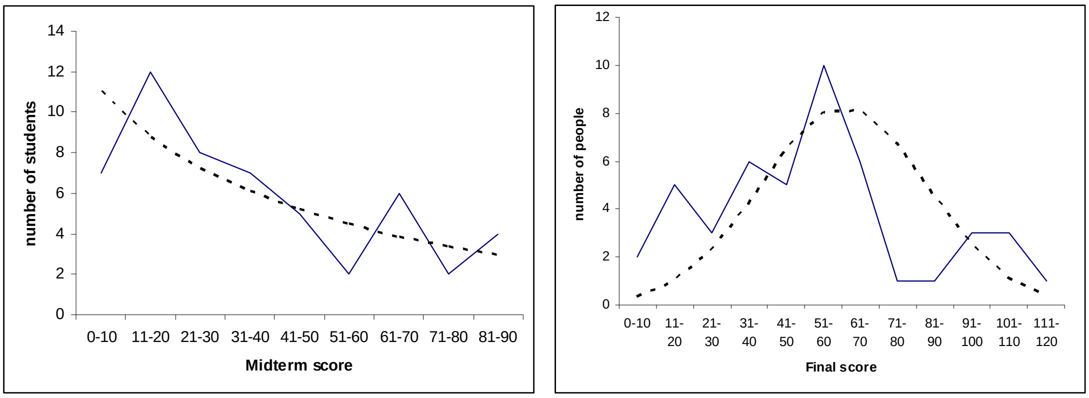
We expected to find an error width that is smaller than the number of ability categories (maximum width). Table 3 shows that the best-fit criterion requires us to use in the model the widest possible error (over all ability categories). The sum of absolute differences between the simulated and experimental perceived score distribution was steadily (even though slowly) decreasing when we were increasing the error width. The results of the model suggest that either people are making big errors in judgment or that there are features that our simple model lacks. Apart from the result of the calibrated error computations, we will discuss the distributional assumption of the model and the model's ability to replicate the pattern generated by the experiment. We will first look at the real distributions.

Figure 3. *The distribution of people over real score (solid line) in the midterm exam approximated with a J-distribution and in the final exam approximated with a normal distribution (both distributions grouped into intervals of 10 score points).*

[31] Note that our model allows only odd number of errors – real ability ± n = 2n+1.

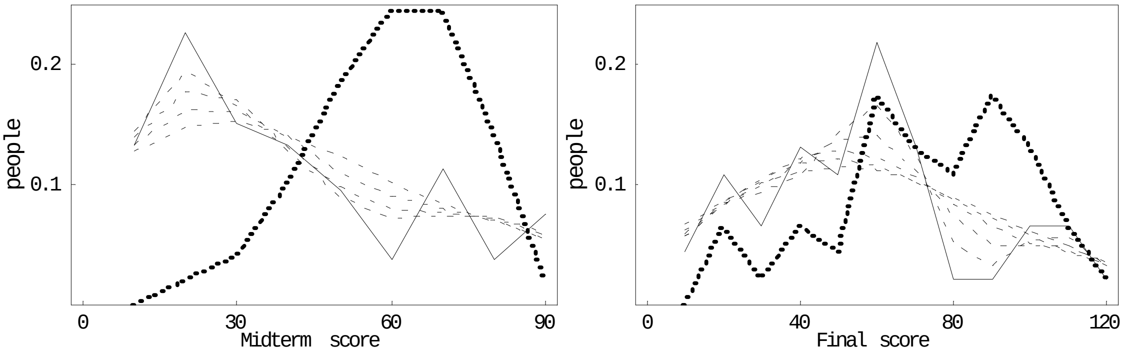
[32] Because of the low number of subjects (46) and high number of ability categories (91) our distribution was not dense enough (contained a lot of empty ability categories). In order to get smoother distribution, we grouped abilities into intervals of 5 categories, thus we got 18 categories (this yields on average over 2 people in 1 new ability category).

[33] For the same reason, we grouped abilities into intervals of 10 categories, thus we got 9 categories (this yields on average over 5 people in 1 new ability category).



Note that in the case of the midterm exam, we get a distribution very similar to a J-distribution with a somewhat smoother drop-off at the lower end – similar to what Krajč and Ortmann (2008) conjectured to be relevant in the Cornell and Chicago scenarios because of legacy cases and that they therefore address in their robustness tests. However, the results of the final exam look more normal-distributed with slightly more people in the lower half (and thus serve only as weak support for the asymmetry assumption of the model).

Figure 4. *The distribution of people over real score (solid line), experimental perceived score (dotted line), and simulated perceived score with various error widths (dashed line – flatter curves correspond to results with wider errors) in the midterm and the final exam (both grouped into intervals of 10 score points).*



Does the model generate patterns similar to the experiment? In the first graph we see that the results generated by the model are very different from the experimental results. There is slight overestimation (very small underestimation) among the unskilled (skilled) in the simulated data compared to huge overestimation (much higher underestimation) in the experimental data. Note that the model replicates the asymmetry. The fit (of simulated to experimental data) in the second graph is a little bit better but it seems that this is only due to better calibration of our subjects. Distributions with a zigzag shape close to boundaries cause problems for our model (as it is in the left part of the Final score distribution and right part of the Midterm score distribution).[34] As the assessment process is very complicated, there might be something hidden that the simple model does not capture. We will discuss the reasons for the unimpressive performance of the model after reporting the results from Experiments 2 and 3.

5.1.2. Experiments 2 and 3

Table 4 summarizes the results of simulations informed by the model proposed in Krajč and Ortmann (2008) (steps 1 through 4: Create a distribution of real score from the experimental data, apply the model, compute the "simulated perceived score distribution" and the sum of absolute differences for every possible error width, identify the "calibrated error") with real distributions from tasks in Experiments 2 and 3.

Table 4. *Model performance.*

	# of categories*	Error width (categories)**	Average deviation***
Experiment 2 Stage 1	18	9	0.48
Experiment 2 Stage 2	19	3	0.57
Experiment 3 Stage 1	14	13	0.46
Experiment 3 Stage 2	14	13	0.84

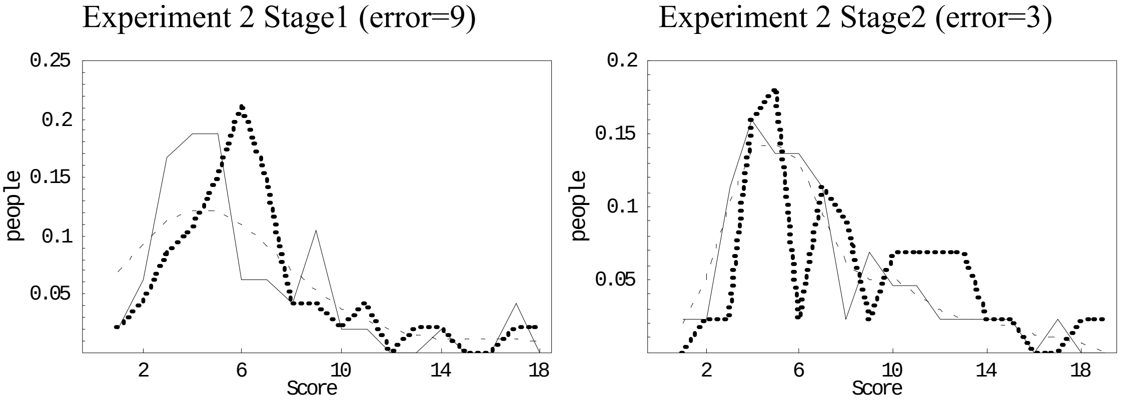
* - The number of ability categories over which people were distributed
** - Width of the error where the model fit is the best
*** - Minimum sum of absolute differences (best fit)

Table 4 suggests that the model seems to work according to our expectations Experiment 2 is where the best fit happens in both stages, with an error smaller than the whole range of abilities (unlike in Experiments 1 and 3). We observe that the best fit requires the model to use an error distributed over 9 (out of 18) categories in Stage 1. This error distribution narrows in Stage 2 to 3 categories.[35] This suggests that there is a positive effect of time on calibration – people commit fewer errors when they do the task for the second time. Figure 5 depicts these results. In Experiment 3 (similar to Experiment 1), however, the best fit gave us the widest possible error.

[34] The reason probably is that in such a case our model mostly equalizes (averages) the oscillation of the real score distribution.

[35] The average achieved score in Experiment 2 increased from 6.71 in Stage 1 to 7.53 in Stage 2; the average score achieved among those who did both tasks was 7.1 in Stage 1 and 8.08 in Stage 2. Since the average score increased, the increase in miscalibration was not caused by the change in average score. The experiment in Brueggen and Strobel (2008) demonstrated a similar result – utility in chosen-effort tasks is similar to utility in real-effort tasks.

Figure 5. *The distribution of people over real score (solid line), experimental perceived score (dotted line), and simulated perceived score (dashed line) in Experiment 2 Stage 1 and Stage 2.*

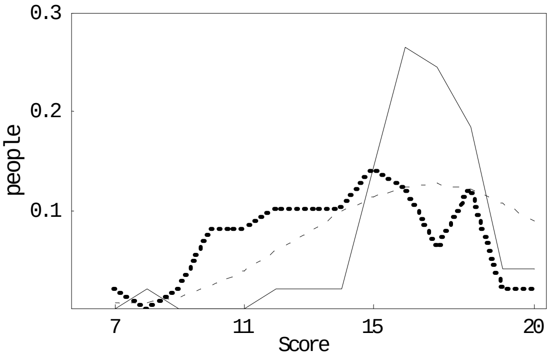


Experiment 2 Stage1 (error=9)

Experiment 2 Stage2 (error=3)

Despite the fact that in Experiment 3, similar to Experiment 1, the calibrated error equals to the highest possible number, the following graph shows that the model works quite well, in Experiment 3, although the score distribution is not J-distributed (more people with lower abilities, but also more people with good abilities).

Figure 6. *The distribution of people over real score (solid line), experimental perceived score (dotted line), and simulated perceived score (dashed line) in Experiment 3 Stage 1.*



Recall that in this case people exhibited, on average, underconfident behavior (mean= -2.28). From the graph we clearly see that the perceived abilities distribution generated by the model is very similar to the one generated by the data. Note that there is a significant shift of people from higher ability categories towards lower ones (even though it is not true for a couple of the very top categories). Thus, in this case the model works for the reversed J-distribution of real abilities even though in this situation the best fit for Own score is given by the error of maximum width. For Stage 2 we get very similar results.

In sum, the simulations of our model show that although it does to some extent capture the patterns observed in the experimental data, it does not fit the experimental data particularly well. The main reason for the unimpressive performance of the model is the skill distribution of the subject pool used in the experiments. The model was originally designed on the distribution of skills one typically finds in the subject pool consisting of Cornell and Chicago university students. While CERGE-EI prep students are sampled from the upper half of the relevant population due to various policy programs (e.g., to support students from certain developing countries), they are on average not sampled from the same part of the abilities distribution as U.S. students at top programs such as Cornell and Chicago.[36] In addition, the motivational incentives are much higher for Cornell and Chicago students. We therefore cannot justify the assumption of the J-distribution of skills which the model assumes (and for that we did not find perfect support in the data).[37] However, it also is possible that the model lacks some important feature that is present in the assessment process. Nevertheless, the model might also work for other distributions (as we have shown for reversed J-distribution).

5.2a. Hypothesis 2a - general information

General information decreases miscalibration.

In each subsection, we will focus our analysis on Own score and Percentile predictions.

5.2a.1. Experiment 1

5.2a.1.1. Descriptive results

Own. Table 5 summarizes the basic statistics of miscalibration in Own score as well as of the results of exams.[38] Note that Mean and St. Dev. Own denotes miscalibration of people in Own score predictions while Mean score and St. Dev score denotes actual score (out of 90) achieved on the particular exam.

[36] In addition, our experiments were conducted with students in the admission process while the model's assumptions are based on the distribution of skills of regular Cornell and Chicago University students.

[37] Note that the model is not restricted to strict J-distribution. For example, for a uniform distribution it generates overestimation of the unskilled equal to underestimation of the skilled.

[38] In order to have comparable numbers all values were computed with adjusted data (as explained above).

Table 5. *Miscalibration in Own score and exam results.*

Own score	Midterm (prediction 1)	Midterm (prediction 2)	Final
Mean Own	30.07	26.30	12.85
St. Dev. Own	23.21	20.20	15.25
Mean score	36.47		39.85
St. Dev. score	25.12		21.63

The average miscalibration of Own score predictions in midterm and final in Experiment 1 decreased over time [30.07→26.30→12.85]; so were standard deviations and hence stability of calibration [23.21→20.20→15.25]. We also see that the difficulty of both exams was approximately the same (a bit lower in the final exam [36.47→39.85]). The variance of students' scores was a little bit lower in the final exam [25.12→21.63].

Percentile. Similarly, the average miscalibration of Percentile predictions in midterm and final in Experiment 1 decreased over time [0.23→0.20→0.11][39]; standard deviations (stability of calibration) decreased only after the midterm exam [0.27→0.28→0.22].

Recall that in the case of exam predictions, students received direct feedback after the midterm. However, we observe improved calibration in all three measures already before this information was revealed – this is most likely based on indirect feedback obtained from interactions with classmates. The improved calibration is reflected in a shift in the perceived Own score curve (which might have been caused by lowering the expectation of the average score or of difficulty) and the rotation of the trend line (increase in its slope), as also shown in the graphs in Appendix A1.

5.2a.1.2. Statistical results

In order to test our hypotheses, we implemented two approaches. First we tested for the significance of the difference between real score and estimated score distributions. Second, we tested for equality of error distributions (=miscalibration) between the predictions. Thus, in the first case we tested whether miscalibration at a point of time is

[39] These numbers can also be found in Table 2a.

significant and in the second case, whether there is improvement in calibration over time. When our samples were dependent (repeated measurements on a single sample), we used the Wilcoxon matched-pairs signed-ranks test[40] (Wilcoxon signed-rank test – WSR) to test whether two samples of observations[41] come from the same distribution. In case the samples were independent, we used the Mann-Whitney-Wilcoxon test (MWW).[42]

In addition to statistical tests, we also computed the strength of the effect (e.g. of information or time) – effect sizes.[43] Since our samples are in some cases not big enough, effect sizes might help us suggest a relation between the variables under investigation. Concretely, we computed Cohen's d[44] that measures the effect size (of information or time) when used on errors of predictions from two different time spots.

<u>Own.</u> First, using the WSR test, we tested for the significance of the difference between real score and estimated score distributions in Experiment 1 for Own score estimates.

Table 6. *The Wilcoxon signed-rank test on real and estimated Own score.*

Real vs. estimated	M1	M2	F
p-value	**0.0000**	**0.0000**	**0.0000**

The WSR test rejected equality of distributions (real Own score and predicted Own score) in all cases at the 1% significance level. These results suggest that the predictions were very inaccurate.

Second, we tested for equality of miscalibration between exam predictions. To determine how strong the effect of information was, we also computed Cohen's d.

[40] Wilcoxon matched-pairs signed-ranks is a non-parametric test that tests the equality of matched pairs of observations. The null hypothesis is equality of distributions.

[41] We included only those participants whose data were available in both samples under investigation.

[42] Null hypothesis in the Mann-Whitney-Wilcoxon test is equality of distributions. The null hypothesis is that the two samples are drawn from a single population, and therefore that their probability distributions are equal.

[43] Effect size measures the strength of the relationship between two variables. Effect size measures are often used to determine the importance of the relationship when there are not enough observations to reach statistical significance.

[44] Cohen's d measures the effect size on means. It is computed as the difference between means of the two distributions divided by the pooled standard deviation. One could compute the effect size using the paired t-test rather than the original pooled standard deviations from the two means. However, Dunlop et al (1996) argue that in such a case the effect size would overestimate the real effect size. Therefore, we used the conventional way of computing Cohen's d.

Table 7. *The Wilcoxon signed-rank test on errors (miscalibration) in Own score and Cohen's d (effect size and effect intensity).*

Errors	M1M2	M2F	M1F
p-value	0.2604	**0.0008**	**0.0018**
Cohen's d	-0.17	-0.75	-0.88
Effect size[45]	small	large	large

While the WSR test did not reject equality of distributions of errors from midterm predictions 1 and 2, equality of error distributions from midterm predictions 1 and 2 and final prediction can be rejected at the 1% significance level. The effect size computations are in line with statistical results. These results, together with the observation that means of miscalibration are decreasing over time, support our Hypothesis 2a that there is less miscalibration with more information.

We also tested the above mentioned slope of the trend line of predicted Own score in midterm prediction 1. We regressed the predicted score on constant and trend; both were significant (p-values=0.000, 0.038, respectively). Thus, we can conclude that people do not have completely random prior beliefs about their absolute performance.

In sum, we identified statistically significant miscalibration in all three predictions about Own score; yet we observed a significant improvement in calibration over time.

Percentile. First, we used the WSR test to test the difference between real percentile and estimated percentile distributions in Experiment 1.

Table 8. *The Wilcoxon signed-rank test on real and estimated Percentile.*

Real vs. estimated	M1	M2	F
p-value	**0.0000**	**0.0000**	**0.0010**

The WSR test rejected equality of distributions in the case of both midterm predictions and final prediction at the 1% significance level, indicating significant miscalibration.

Second, we tested for equality of error distributions between exam predictions.

[45] According to Cohen's classification: 0.2 = small, 0.5 = medium, 0.8 = large effect.

Table 9. *The Wilcoxon signed-rank test on errors (miscalibration) in Percentile and Cohen's d (effect size and effect intensity).*

Errors	M1M2	M2F	M1F
p-value	0.1328	0.3581	0.1158
Cohen's d	-0.1	-0.35	-0.45
Effect size	small	medium	medium

The WSR test was not able to reject equality of distributions of errors from any combination of midterm predictions 1 and 2 and final prediction. However, we were close to rejecting equality of error distributions from midterm prediction 1 and final prediction (p-values=0.1158) at the 10% significance level. Even though we did not reach significance in statistical tests, the effect sizes are medium between any of midterm predictions and final prediction (higher for the midterm predictions 1 and final prediction).

We also tested the slope of the trend line of predicted Percentile in midterm prediction 1. We regressed the predicted Percentile on constant and trend; both were significant (p-values=0.000, 0.090, respectively). Thus we conclude that like the case of Own score predictions, people do not have completely random prior beliefs about their relative rank position.

In short, we identified statistically significant miscalibration in Percentile. We did not find statistically significant improvement of calibration over time but the effect sizes suggest a medium effect of time on calibration.

5.2a.2. Experiments 2 and 3

5.2a.2.1 Descriptive results

Own. The following table summarizes the basic statistics of miscalibration in Own score as well as of the results of tasks.[46]

[46] In order to have comparable numbers all values were computed with adjusted data (as explained above).

Table 10. *Task results and miscalibration in Own score.*[47]

Own score	Experiment 2			Experiment 3		
	Stage 1	Stage 2	Stage 2 (NF)	Stage 1	Stage 2	Stage 2 (NF)
Mean Own	0.62	1.45	1.87	-2.28	0.39	0.73
St. Dev. Own	1.65	2.25	2.53	2.68	3.19	3.43
Mean score	6.85	7.70	7.70	16.41	12.71	12.31
St. Dev. score	3.55	3.63	4.29	1.99	2.32	2.43

Looking at the basic results in Table 10 we can see that the mean of miscalibration increased for estimates of Own score in Experiment 2 [0.62→1.45] but decreased in Experiment 3 [-2.28→0.39]; the standard deviation increased in both tasks [1.65→2.25; 2.68→3.19]. We also can see that while in Experiment 2 students correctly solved, on average, more problems in Stage 2 than in Stage 1, the reverse holds for Experiment 3. This table also suggests that the stability of calibration is not improving in mathematical skill task and general-knowledge tasks. However, calibration was, on average, reasonably good already in Stage 1. Note that unlike Experiment 1, students are in Experiment 2 and 3 pretty well calibrated in Own score estimates already in Stage 1 (see the graphs in Appendix A2).[48]

Percentile. The Percentile estimates were on average more accurate in Experiment 2 [0.11→0.09][49] than in Experiment 3 [0.15→0.11]; overestimation decreased over time in both tasks. Standard deviation almost did not change in Experiment 2 [0.30→0.28] but increased in Experiment 3 [0.30→0.36].[50] Note that the Percentile predictions, unlike the Own predictions, improved in Stage 2 in both tasks.

5.2a.2.2. Statistical results

[47] We also computed the basic statistic separately for those who did not get additional feedback in Stage 2 (denoted as NF in the table) in order to separate the effect of additional feedback that is investigated in hypothesis 2b.

[48] We also computed the number of people whose calibration in Own score estimates improved/did not change/worsened in Stage 2 (compared to Stage 1). The results can be found in Appendix A3.

[49] These numbers can also be found in Tables 2b and 2c.

[50] Using the power computations, we computed the sample size needed to get 95% statistical significance with 80% power of the test. In our results, the difference between means between stages is above 1 and the standard deviation is above 2, resulting into medium effect size = 0.5. Thus, to get the desired significance level, we would need 33 subjects. So, our sample size is big enough (if we include all participants).

Own. In order to test Hypothesis 2a we first tested the difference between real score and estimated score distributions in both tasks and both stages for Own score estimates.

Table 11. *The Wilcoxon signed-rank test on real and estimated Own score.*

Real vs. estimated	Experiment 2			Experiment 3		
	Stage 1	Stage 2	Stage 2 (NF)	Stage 1	Stage 2	Stage 2 (NF)
p-value	**0.0049**	**0.0002**	**0.0063**	**0.0000**	0.5537	0.4543

We rejected equality of distributions in both stages of Experiment 2 and in Stage 1 of Experiment 3 at the 1% significance level. We were not able to reject the null hypothesis in Stage 2 of Experiment 3, suggesting good calibration. These results indicate that our Hypothesis 2a is supported only in Experiment 3 because in Stage 1, the distributions of real and estimated score were significantly different while in Stage 2 they were not. Note that miscalibration in Experiment 3 was insignificant already in Stage 2 while in Experiment 1 it was significant even after 3 iterations (estimates in time). We will discuss these observations in more detail in the discussion section.

Second, we used the WSR test to test for equality of error distributions between stages. Note that the mean error in Own score estimates in Experiment 3 is negative in Stage 1 and positive in Stage 2. Thus, analyzing these data, we would investigate the significance of increase in overconfidence (or decrease in underconfidence). However, we are more interested in the distance of the mean miscalibration from zero.[51] Therefore we did the same tests for Experiment 3 but we transformed the sign of all data from Stage 1 (which is equivalent with reversion of the sign of the mean). We did this transformation always when there was a positive mean.

Table 12. *The Wilcoxon signed-rank test on errors in Own score and Cohen's d.*

Errors	Experiment 2	Experiment 2 (NF)	Experiment 3	Experiment 3 (|mean|)	Experiment 3 (NF)	Experiment 3 (NF,|mean|)
p-value	0.1412	0.3390	**0.0000**	**0.0192**	**0.0021**	0.4336
Cohen's d	0.43	0.59	0.91	-0.70	0.98	-0.50
Effect size	medium	medium	large	large	large	medium

We were not able to reject equality of distributions in Experiment 2, yet we were able to do so in Experiment 3 at the 1% significance level. These results were confirmed in the analysis of subjects without additional feedback as well as in the analysis with positive means. The effect size computations further support our statistical findings.

[51] Specifically, is 0.39 (0.79 for NF subjects) significantly lower than 2.28 (with the corresponding standard deviations)?

The results therefore suggest that the improvement in calibration in Own score was significant in Experiment 3 and that the deterioration of calibration in Experiment 2 was not significant.

<u>Percentile.</u> In order to test Hypothesis 2a we first tested the difference between real Percentile and estimated Percentile distributions in both tasks in Stage 1 and Stage 2.

Table 13. *The Wilcoxon signed-rank test on real and estimated Percentile.*

Real vs. estimated	Experiment 2			Experiment 3		
	Stage 1	Stage 2	Stage 2 (NF)	Stage 1	Stage 2	Stage 2 (NF)
p-value	**0.0249**	0.1923	0.2273	**0.0039**	**0.0806**	**0.0228**

We rejected equality of distributions in Stage 1 in both experiments at the 5% and 1% significance level, respectively, and in Stage 2 of Experiment 3 at the 10% significance level. We were not able to reject equality of distributions in Stage 2 of Experiment 2. The results were qualitatively the same when we included only those without additional feedback. These results suggest that our Hypothesis 2a is supported in Experiment 2 where miscalibration turned out to be non-significant in Stage 2 but not in Experiment 3 even though the significance level of rejecting equality of distributions increased in Experiment 3 from 1% to 10%.

Second, we used the WSR test to test for equality of error distributions between stages.

Table 14. *The Wilcoxon signed-rank test on errors in Percentile and Cohen's d.*

Errors	Experiment 2	Experiment 2 (NF)	Experiment 3	Experiment 3 (NF)
p-value	0.6401	0.8248	0.6352	0.9405
Cohen's d	-0.06	0.16	-0.1	0.16
Effect size	none	small	small	small

We were able to reject equality of distributions neither in Experiment 2 nor in Experiment 3. The effect size computations identified at most small effect sizes.

We thus found weak support for Hypothesis 2a only in Experiment 2 where miscalibration turned out not to be statistically significant in Stage 2. However, the improvement in calibration was not identified as statistically significant.

5.2b. Hypothesis 2b – feedback

Lower miscalibration with specific information (feedback) than without it.

5.2b.1. Experiment 1

Not applicable.

5.2b.2. Experiments 2 and 3

To recall, in Stage 2 we gave full feedback from Stage 1 to approximately half of our subjects. Those with feedback were informed about their own score, the percentage of better performing people, and the average score in that particular task in Stage 1.

5.2b.2.1. Descriptive results

Own. Table 15 displays the average overconfidence and standard deviations of miscalibration of subjects with and without feedback in Stage 2.[52]

Table 15. *Miscalibration in Own score in the feedback and non-feedback treatment.*

Experiment 2	Own (feedback)	Own (no feedback)
# of subjects	21	23 (19)[53]
Mean	1.00	1.87 (1.84)
St. Dev.	1.84	2.53 (2.59)
Experiment 3	Own (feedback)	Own (no feedback)
# of subjects	19	26 (21)
Mean	-0.08	0.73 (0.76)
St. Dev.	2.86	3.43 (3.39)

Table 15 shows that both measures of miscalibration – mean and standard deviation – are in both tasks lower in the feedback treatment than in the non-feedback treatment; this is in line with the prediction of our Hypothesis 2b.[54]

[52] We included in this analysis also those people who did not participate in Stage 1. We computed these numbers also without them and the results did not differ much and did not change qualitatively.

[53] The first number is computed including all subjects who participated in Stage 2. In parenthesis, we report statistics computed including only those subjects who also participated in Stage 1.

[54] We also counted the number of better/worse/equally good performing people. Results can be found in Appendix A4.

Percentile. Table 16 displays the average overconfidence and standard of miscalibration of subjects with and without feedback in Stage 2.[55]

Table 16. *Miscalibration in Percentile in the feedback and non-feedback treatment.*

Experiment 2	Percentile (feedback)	Percentile (no feedback)
# of subjects	21	23 (19)[56]
Mean	0.02	0.16 (0.09)
St. Dev.	0.15	0.34 (0.31)
Experiment 3	Percentile (feedback)	Percentile (no feedback)
# of subjects	19	26 (21)
Mean	0.00	0.20 (0.20)
St. Dev.	0.36	0.34 (0.33)

Table 16 shows that mean overconfidence is almost zero in both tasks for people with feedback and positive for the others; this is in line with our Hypothesis 2b. Standard deviation is smaller in the feedback treatment in Experiment 2 yet the same in Experiment 3. In addition, the results suggest that, on average, subjects with feedback in Stage 2 outperform (in calibration) all subjects in Stage 1 in both tasks.

5.2b.2.2. Statistical results

Own. First we tested for the difference in Own score estimate and actual Own score of those with and without feedback. As our samples are correlated in this case we used the WSR test. We were able to reject equality of distributions neither for feedback nor for non-feedback treatment of Experiment 2 (p-value=0.2643, 0.3390, respectively) – good calibration in both treatments. We rejected equality of distributions for both treatments in Experiment 3 (p-value=0.0048, 0.0021) – significant miscalibration in both treatments.

Second, to test whether the two samples of miscalibration (with feedback and without feedback) come from the same distribution we used the MWW test. However, we were not able to reject the null hypothesis (equality of distributions) in any of the tasks. Effect size is medium in Experiment 2 and small in Experiment 3.

[55] We included in this analysis also those people who did not participate in Stage 1. We computed these numbers also without them and the results did not differ much and did not change qualitatively.

[56] The first number is computed including all subjects who participated in Stage 2. In parenthesis, we report statistics computed including only those subjects who also participated in Stage 1.

Table 17. *Cohen's d.*

	Experiment 2	Experiment 3
p-value	0.4838	0.5377
Cohen's d	0.39	0.26
Effect size	medium	small

Although the difference between the feedback and non-feedback subjects seems to be substantial, it was not captured by significance tests (probably because of small number of observations). Yet Cohen's d shows some support for our Hypothesis 2b.

Percentile. First, we tested for the difference in Own score estimate and actual Own score of those with and without feedback. The WSR test was not able to reject the null hypothesis in any treatment of Experiments 2 and 3 (p-values=0.6274, 0.8248, 0.5732, 0.9405) indicating statistically non-significant miscalibration.

Second, with the MWW test, we were able to reject equality of error distributions from feedback and non-feedback treatment in Experiments 2 and 3 at the 5% and 10% significance levels, respectively. The results of effect sizes support these statistical results.

Table 18. *Cohen's d.*

	Experiment 2	Experiment 3
p-value	**0.0484**	**0.0848**
Cohen's d	0.54 (0.26)[57]	0.56 (0.56)
Effect size	medium (small)	medium (medium)

In brief, we identified a statistically significant impact of full feedback on calibration in Percentile which was supported with medium effect sizes in both Experiments.

5.3. Hypothesis 3 – Own score vs. Percentile

There is less miscalibration in Own score estimates than in Percentile estimates.

5.3.1. Experiment 1

5.3.1.1. Descriptive results

[57] In parenthesis is Cohen's d computed only with those subjects who also participated in Stage 1.

We first expressed miscalibration of Own score predictions in percentage[58] in order to do the required comparison. The transformed results from Experiment 1 suggest that calibration in estimating Own score is lower than in estimating Percentile in all three cases. However we could say that, in time, miscalibration is declining in both estimates and is also stabilizing.

Table 19. *Adjusted miscalibration in Own score and Percentile.*

Midterm prediction 1	Own	Percentile
Mean	0.33	0.23
St. Dev.	0.26	0.27
Midterm prediction 2	Own	Percentile
Mean	0.29	0.20
St. Dev.	0.22	0.28
Final prediction	Own	Percentile
Mean	0.14	0.11
St. Dev.	0.17	0.22

5.3.1.2. Statistical results

We used the WSR test to test the difference between errors from Own score predictions and Percentile predictions.

Table 20. *The Wilcoxon signed-rank test on errors from Own score and Percentile and Cohen's d.*

	Midterm prediction 1	Midterm prediction 2	Final prediction
p-value	**0.0001**	**0.0007**	**0.0294**
Cohen's d	0.38	0.36	0.15
Effect size	medium	medium	small

We were able to reject equality of distributions of errors in midterm predictions 1 and 2 at the 1% significance level and in final prediction at the 5% significance level. These results show that the difference in these two types of calibration is significant in all three predictions. Remember that our hypothesis is supported in the reverse direction: Percentile predictions are more accurate than Own score predictions. The effect sizes are in line with the statistical predictions.

5.3.2. Experiments 2 and 3

5.3.2.1. Descriptive results

[58] We divided the midterm and final results by 90.

Visual inspection of the adjusted results in the table below suggests that miscalibration is, on average, much lower in estimation of Own score than in estimation of Percentile. The same holds for standard deviations. These results seem to be robust in both experiments and both stages.

Table 21. *Adjusted miscalibration in Own score and Percentile.*

	Stage 1		Stage 2	
Experiment 2	Own	Percentile	Own	Percentile
Mean	0.03	0.11	0.07	0.09
St. Dev.	0.08	0.30	0.11	0.28
Experiment 3	Own	Percentile	Own	Percentile
Mean	-0.11	0.15	0.02	0.11
St. Dev.	0.13	0.30	0.16	0.36

5.3.2.2. Statistical results

We used the same test (WSR test) as in Experiment 1 for testing for the difference in Own score and Percentile estimates.[59]

Table 22. *The Wilcoxon signed-rank test on errors from Own score and Percentile.*

	Experiment 2		Experiment 3	
	Stage 1	Stage 2	Stage 1 (|mean|)	Stage 2
p-value	0.1095	0.6671	**0.0000** (0.6518)	0.1190
Cohen's d	0.36	-0.09	0.17	0.32
Effect size	medium	none	small	small

We rejected the null hypothesis at the 1% significance level only in Stage 1 of Experiment 3; however, when comparing only the absolute miscalibration, these two distributions were not significantly different. Note that we were close to rejecting equality at the 10% significance level also in Stage 1 of Experiment 2 and Stage 2 of Experiment 3. Nor effect sizes support our hypothesis.

Therefore we conclude that the difference in Own score and Percentile miscalibration is not statistically significant in Experiments 2 and 3 even though the difference is close to significant at the 10% significance level in some cases.

5.4. Hypothesis 4 – skills vs. general knowledge

Skill-oriented tasks generate less miscalibration than general knowledge-oriented tasks.

[59] As in testing the Hypothesis 2a we used absolute values of all means. Here, it affects only the mean in Stage 1 of Experiment 3.

5.4.1. Experiment 1

Not applicable.

5.4.2. Experiments 2 and 3

5.4.2.1. Descriptive results

Own. In order to test this hypothesis we compared the results from Experiments 2 and 3. In both stages, the error of Own score estimates is less variable in Experiment 2 than in Experiment 3. However, we cannot draw any conclusion from the results on calibration. We can to some extent explain the variability of error. In the summing problems task students solved a number of problems (majority much less than 20) while in the general-knowledge task the vast majority of our subjects answered all 20 questions. It seems self-evident that there is more space for error in a 20 questions-estimate than in a 9 questions-estimate (9 was the average number of answered problems in Experiment 2).

Table 23. *Basic statistics of errors in Own score.*

Own score	Stage 1		Stage 2	
	Experiment 2	Experiment 3	Experiment 2	Experiment 3
Mean	0.62	-2.28	1.45	-0.39
St. Dev.	1.65	2.68	2.25	3.19

Percentile. The basic statistics for Percentile estimates suggest that our subjects were better calibrated in Experiment 2 than in Experiment 3.

Table 24. *Basic statistics of errors in Percentile.*

Percentile	Stage 1		Stage 2	
	Experiment 2	Experiment 3	Experiment 2	Experiment 3
Mean	0.11	0.15	0.09	0.11
St. Dev.	0.30	0.30	0.28	0.36

5.4.2.1. Statistical results

Significance tests

Own. We used the WSR test to determine if there is statistical significance between errors from Experiments 2 and 3 in each stage. We again used absolute values of means where necessary.

Table 25. *The Wilcoxon signed-rank test on errors in Own score and Cohen's d.*

Own score	Stage 1	Stage 2
p-value	**0.0000**	0.1353
Cohen's d	-0.75	0.39
Effect size	large	medium

We were able to reject equality of error distributions in Stage 1 at the 1% significance level and thus conclude that our subjects were better calibrated in the skill-oriented task. In Stage 2, we were not able to reject the null hypothesis. Effect sizes are in line with our statistical results (it is smaller in Stage 2).

Percentile. We did the WSR test also for Percentile estimates.

Table 26. *The Wilcoxon signed-rank test on errors in Percentile and Cohen's d.*

Percentile	Stage 1	Stage 2
p-value	0.6798	0.3399
Cohen's d	-0.12	-0.07
Effect size	none	none

In the case of Percentile estimates we were not able to reject equality of error distributions in Experiments 2 and 3 in any stage. Nor do effect size computations show any effect. Therefore, we conclude that there is almost no difference in (mis)calibration of relative standing in skill-oriented tasks and general knowledge-oriented tasks.

6. Discussion and conclusion

The results of the analysis of Own score are summarized in Table 27a and the results of the analysis of Percentile are summarized in Table 27b.

Table 27a. *Results of* ***Own score*** *analyses.*

Hypotheses	Significance	Effect size*	Significance		Effect size*	
	Experiment 1		**Exper.2**	**Exper.3**	**Exper.2**	**Exper.3**
H1: ***The model***	–	–	–		–	
H2a: ***General information***	**S**	**L**	NS**	**S**	M**	**L**
H2b: ***Feedback***	–	–	**WS**	**NS**	**M**	S
H3: ***Own score vs. Percentile***	S**	M**	St.1:**WS** St.2:NS	St.1:NS St.2:**WS**	St.1:**M** St.2:N	St.1:S St.2:S
H4: ***Skills vs. general knowledge***	–	–	Stage 1: **S** Stage 2: **WS**		Stage 1: **L** Stage 2: **M**	

Table 27b. *Results of* ***Percentile*** *analyses.*

Hypotheses	Significance	Effect size*	Significance		Effect size*	
	Experiment 1		**Exper.2**	**Exper.3**	**Exper.2**	**Exper.3**
H1: ***The model***	–	–	–		–	
H2a: ***General information***	NS	**M**	**WS**	NS	S	S
H2b: ***Feedback***	–	–	**S**	**S**	**M**	**M**
H3: ***Own score vs. Percentile***	S**	M**	St.1:**WS** St.2:NS	St.1:NS St.2:**WS**	St.1:**M** St.2:N	St.1:S St.2:S
H4: ***Skills vs. general knowledge***	–	–	Stage 1: NS Stage 2: NS		Stage 1: N Stage 2: N	

* - Cohen’s d
** - the effect in the opposite direction
“ – “ – not available for that experiment
Significance: NS – not supported, WS – weakly supported, S – supported.
Effect size: N – none, S – small, M – medium, L – large

The key results of the experiments reported in this paper can be summarized as follows:

1. Overconfidence prevails in almost all types of estimates/predictions.
2. General information improves calibration over time, especially in absolute self-assessment in (field) Experiment 1.
3. Specific information (feedback) significantly improves calibration in absolute self-assessment in Experiments 2 and 3; though the basic statistics and effect sizes suggest better calibration of subjects with feedback in all types of estimates in both experiments.
4. Absolute self-assessment is more responsive to information than relative self-assessment.
5. Although the simple model proposed by Krajč and Ortmann (2008) is able, to some extent, to capture main patterns of the unskilled-and-unaware problem, it does not explain the experimental data well. The unimpressive performance of the model might be caused by the differences in the subject pools assumed in the model and empirically found in our experiments.

We conducted three experiments in a natural setting: a preparatory semester for PhD students. This real-world situation provides an opportunity to investigate the impact of information (acquired throughout the semester) on absolute as well as relative self-assessment and to test the presence of the unskilled-and-unaware problem in various tasks and under various conditions. The first experiment was a field experiment

(Experiment 1) where students of the prep semester had to predict their performance on the micro midterm (two times) and final exam (one time). Information was provided in a natural way, in the following 2 forms: natural interaction among members of the prep semester cohort throughout the prep semester (math, micro, macro, exercises, lectures, homework, etc.) and the results of the micro midterm exam before final predictions. We measured calibration in absolute self-evaluation (Own score), relative self-evaluation (Percentile), and group evaluation (Average score). The results revealed prevailing overconfidence in almost all types of predictions. We identified clear improvement in calibration with increasing information over time in this experiment; the highest improvement was achieved in Own score (absolute self-assessment). It is impressive how rapidly our subjects improved their calibration, especially given the information acquisition after the midterm exam. We also showed that although the model (Krajč and Ortmann, 2008) replicates some the patterns identified in the experimental data, it did not fit the experimental data very well. We conclude that the unimpressive performance is caused by use of a subject pool with a different structure than the model was originally designed for, or that the model might lack some important feature.

Two laboratory experiments (Experiments 2 and 3) were embedded into the field experiment. In these two experiments, we controlled for information distribution; concretely, complete feedback about one's own absolute, own relative, and group performance was given only to half of the participants. With these experiments we also investigated the difference between self-assessment in skill-oriented tasks (Experiment 2) and general knowledge-oriented tasks (Experiment 3). We measured the same types of miscalibration as in Experiment 1: Own score, Percentile, and Average score. On average, we found that people overestimate their abilities in Experiment 2 and underestimate their abilities in Experiment 3 (especially in Stage 1). Over time, people improved their calibration in both experiments and all types of calibrations, except for Own and Average score in Experiment 2. Moreover, we identified a positive effect of feedback on calibration in all measured variables in Experiment 2. The performance of the simple model proposed in Krajč and Ortmann (2008) is similar to Experiment 1. It can again be explained by the difference in subject pools.

We found only partial support for our last two hypotheses. Skill-oriented tasks generated significantly lower miscalibration than general knowledge-oriented tasks in Own score only in Stage 1. Moreover, we found more overestimation in Own score than in Percentile only in Experiment 1. Based on this observation and from visual inspection of the results, the miscalibration in relative measure seems to be, from our three experiments, more stable than miscalibration in Own score. This observation came as a surprise and suggests that the absolute miscalibration alone is not a very good explanatory measure of relative self-assessment.

In addition to the above reported results, we also investigated how the distribution of over/underestimation of the average score influences the unskilled-and-unaware problem. In Experiment 1, overestimation of Average score is very similar among the skilled and the unskilled in the case of Midterm predictions. Overestimation of Average score is a little bit better in the top one third than in the bottom two thirds in the case of Final predictions. Thus, the perception of difficulty combined with the quality of the group[60] seems not to be dependent on the skills (exam scores). Similarly, in Experiments 2 and 3 we do not find a dependency of overestimation of average group score on performance. Overestimation of the Average score was approximately equally distributed among our subjects. Even though some part of overestimation of Own score might be caused by the expectations of lower exam difficulty, the perception of Average score seems not to affect the overestimation of Own score by the unskilled and underestimation by the skilled. Thus we conclude that the unskilled-and-unaware problem is not caused by different assessment (expectations) of the group quality/task difficulty.

In order to say more about the evolution of the unskilled-and-unaware problem over time (with increasing information) we also analyzed miscalibration in Own score and Percentile by quartiles.

[60] With the data we have, we cannot separate the prediction of the quality of the group from expectations of the exam/task difficulty.

Table 28. *Overestimation in Own score in Experiment 1 by quartiles.*

OC Own score	Midterm prediction 1	Midterm prediction 2	Final prediction
Bottom quartile	49.98	42.96	19.64
2nd quartile	43.77	32	17.66
3rd quartile	31.23	29.62	15.55
Top quartile	-1.80	2.34	-0.25

Over time, we observe a remarkable improvement in calibration in the bottom three quartiles. In addition, we observe decreasing differences in overestimation between quartiles, which means that the bottom quartiles improve their calibration the most. Finally, for the Final prediction we see that average overestimation of the bottom three quartiles is almost the same while people in the top quartile are, on average, well calibrated.

Table 29. *Overestimation in Percentile in Experiment 1 by quartiles.*

OC Percentile	Midterm prediction 1	Midterm prediction 2	Final prediction
Bottom quartile	0.53	0.56	0.29
2nd quartile	0.37	0.22	0.16
3rd quartile	0.14	0.12	0.09
Top quartile	-0.08	-0.08	-0.08

We observe very similar patterns also in Percentile predictions. Note that the underestimation of the top quartile remains the same over time. Overestimation of all other quartiles (but the bottom quartile between Midterm predictions) improves over time; more in the two bottom quartiles than in the third quartile.

Table 30. *Overestimation in Own score in Experiments 2 and 3 by quartiles.*

OC Own score	Experiment 2		Experiment 3	
	Stage 1	Stage 2	Stage 1	Stage 2
Bottom quartile	1.2	2.45	-1.27	-1.89
2nd quartile	1	0.82	-1.58	-2.13
3rd quartile	0.42	1.36	-3.83	0.38
Top quartile	0.27	1.18	-2.5	-1.67

In Stage 1 of Experiment 2, similar to Experiment 1, going from the bottom to the top quartile we observe decreasing overestimation of Own score. However, this pattern is present neither in Stage 2 of Experiment 2 nor in Experiment 3. We cannot draw uniform conclusions about the relationship of the level of overestimation and quartiles.

Table 31. *Overestimation in Percentile in Experiments 2 and 3 by quartiles.*[61]

OC Percentile	Experiment 2		Experiment 3	
	Stage 1	Stage 2	Stage 1	Stage 2
Bottom quartile	0.37	0.33	0.52	0.53
2nd quartile	0.25	0.20	0.24	0.27
3rd quartile	0.03	-0.07	0.02	-0.07
Top quartile	-0.18	-0.08	-0.14	-0.23

We observe patterns similar to those in Experiment 1 – overestimation decreases with increasing quartiles performance. The only difference is that in Stage 2 we observe underestimation already in the third quartile. There is some improvement in calibration in almost all quartiles in Experiment 2, yet no improvement (even worsening) in calibration in Experiment 3.

The quartiles analysis shows that the improvement in calibration (due to better information) was, not surprisingly, mostly driven by the unskilled, which supports our claim that the unskilled-and-unaware problem can be mitigated by providing sufficient information.

Note that in none of our experiments do we observe anyone with a predicted/estimated percentile rank in the worst 20% of the group. There are several possible explanations. First, since we had a CERGE-EI experimenter, students might not have trusted that we would treat the data confidentially and therefore did not want to send a negative signal about themselves. Second, it is socially very complicated to express such a negative opinion, whatever reason might be. For example, people might have preferences for self-esteem and expressing such a negative opinion would harm their self-esteem (e.g., Koeszegi, 2006).

As already mentioned, the issue of representativeness of stimuli is very important in studies on overconfidence involving two-alternative general-knowledge questions (Juslin et al, 2000). We did not have the chance to influence the representativeness of problems given on the midterm and final exams in Experiment 1. However, we were able to do so in Experiments 2 and 3: by choosing our tasks so that we would be able to

[61] Note that the quartiles results of all three experiments are similar to results in Kruger and Dunning (1999).

control for this (at least to the extent that we could implement random sampling of questions from a known reference class). Since we did not have enough subjects to use this treatment separately in Experiments 2 and 3, we could only compare miscalibration from Experiment 1 with miscalibration from Experiments 2 and 3. We identified higher miscalibration in absolute self-assessment in Experiment 1 than in Experiments 2 and 3. Unfortunately, we cannot say whether this difference was caused by (possible) non-representativeness of stimuli used in Experiment 1, or by the different way of gathering predictions/estimates (before/after task)[62], or by the difference in the tasks alone. To answer this question, one should design an experiment in which it is possible to separate these three effects. For relative self-assessment, we found similar but weaker effect.

Our analysis leaves several questions unanswered, which might be subject to further investigation. Our experiments show that there is faster improvement in calibration in absolute than in relative self-assessment. In order to find out how people create estimates about their absolute and relative performance, it would be useful to know what kind of feedback (absolute, relative, and/or average) helps them to improve calibration in absolute self-assessment and what kind of feedback in relative self-assessment. Based on these results one could better understand what the relation between creating absolute and relative self-assessments is.[63] Moreover, the simple model of ability perception (Krajč and Ortmann, 2008) might be extended.

[62] We might have asked for the predictions in Experiments 2 and 3 before the task was performed in order to reduce the difference between Experiment 1 and Experiments 2 and 3. However, it could initiate speculative behavior in order to win the money promised for the best prediction. Note that this does not happen in Experiment 1, where the incentives for as good performance as possible are much higher (admission to CERGE-EI).

[63] E.g., do people estimate first their Own score and then, based on this estimate, their relative standing?

References

Brüggen A., Strobel M., (2008). Real Effort Versus Chosen Effort in Experiments. *Economics Letters, 96 (2), August,* 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 F.C., Hogarth M.R., (1999). The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework. *Journal of Risk and Uncertainty, 19 (1-3)*, 7-42.

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

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

Dhami K.M., Hertwig R., & Hoffrage U., (2004). The Role of Representative Design in an Ecological Approach to Cognition. *Psychological Bulletin, 130 (6)*, 959–988.

Duffy J., Hopkins E., (2005). Learning, Information, and Sorting in Market Entry Games: Theory and Evidence. *Games and Economic Behavior, 51*, 31–62.

Eckel C.C., Grossman P.J., (2000). Volunteers and Pseudo-Volunteers: The Effect of Recruitment Method in Dictator Experiments. *Experimental Economics, 3 (2),* 107-120.

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

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

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

Erev I., Wallsten T.S., & Budescu D.V., (1994). Simultaneous Over- and Underconfidence: The Role of Error in Judgment Processes. *Psychological Review, 101,* 519-527.

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

Gigerenzer G., Hoffrage U., & Kleinboelting H., (1991). Probabilistic Mental Models: A Brunswikian Theory of Confidence. *Psychological Review, 98 (4)*, 506-528.

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

Hoelzl E., Rustichini A., (2005). Overconfident: Do You Put Your Money on It? *Economic Journal, 115, April*, 305-318.

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.

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

Koeszegi B., (2006). Ego Utility, Overconfidence, and Task Choice. *Journal of the European Economic Association, 4 (4),* 673–707.

Krajč M., Ortmann A., (2008). Are the Unskilled Really That Unaware? An Alternative Explanation. *Journal of Economic Psychology, 29 (5),* 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.

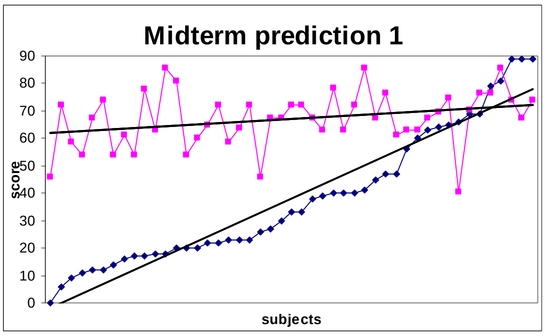
Niederle M., Vesterlund L., (2007). Do Women Shy Away from Competition? Do Men Compete Too Much? *The Quarterly Journal of Economics, 122 (3),* 1067-1101.

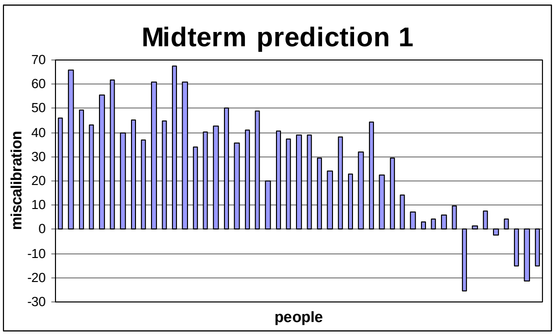
Rydval O., Ortmann A., (2004). How Financial Incentives and Cognitive Abilities Affect Task Performance in Laboratory Settings: An Illustration. *Economics Letters, 85 (3),* 315-320.

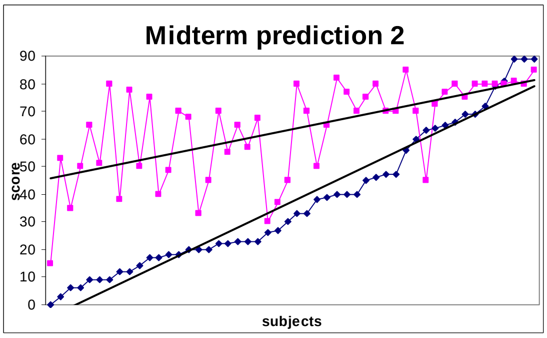
Smith V.L., (2002). Method in Experiment: Rhetoric and Reality. *Experimental Economics, 5 (2),* 91-110.

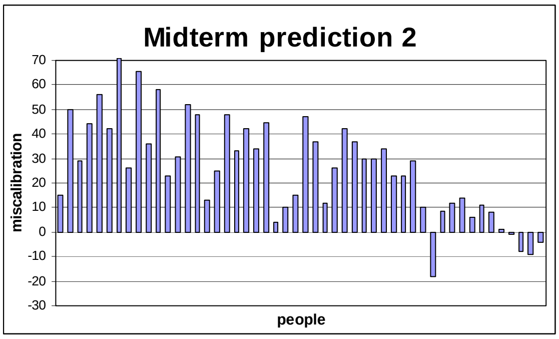
Appendix A1

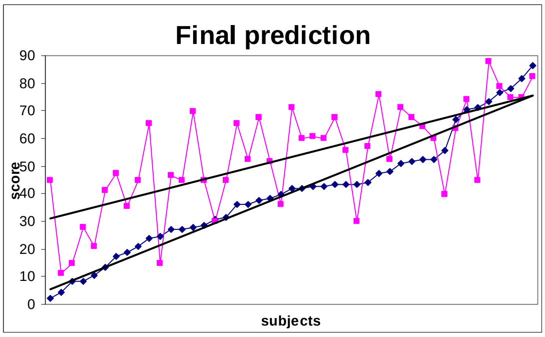
On the left, graphs of the real (blue line) and estimated (pink line) distribution of score (together with the corresponding trend lines). On the right, graphs of distribution of miscalibration of Own score (all data adjusted and ordered from the lowest to the highest real score).

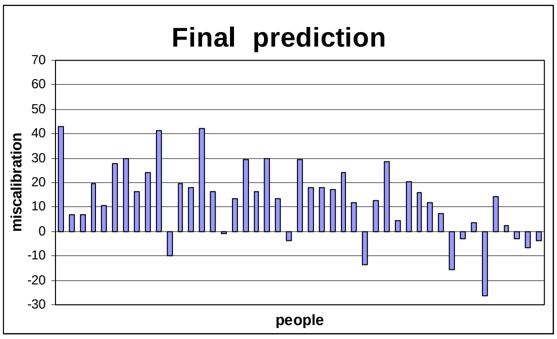




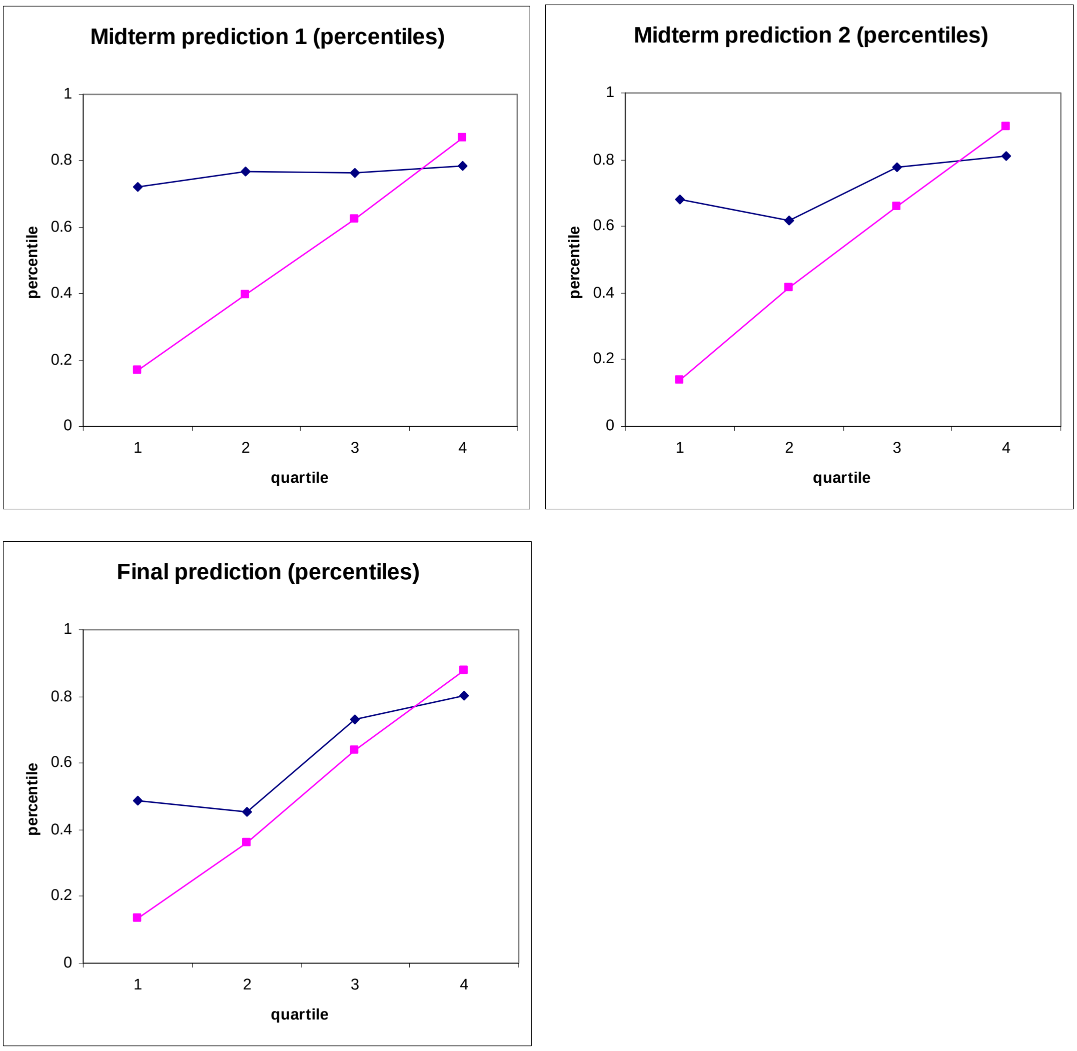






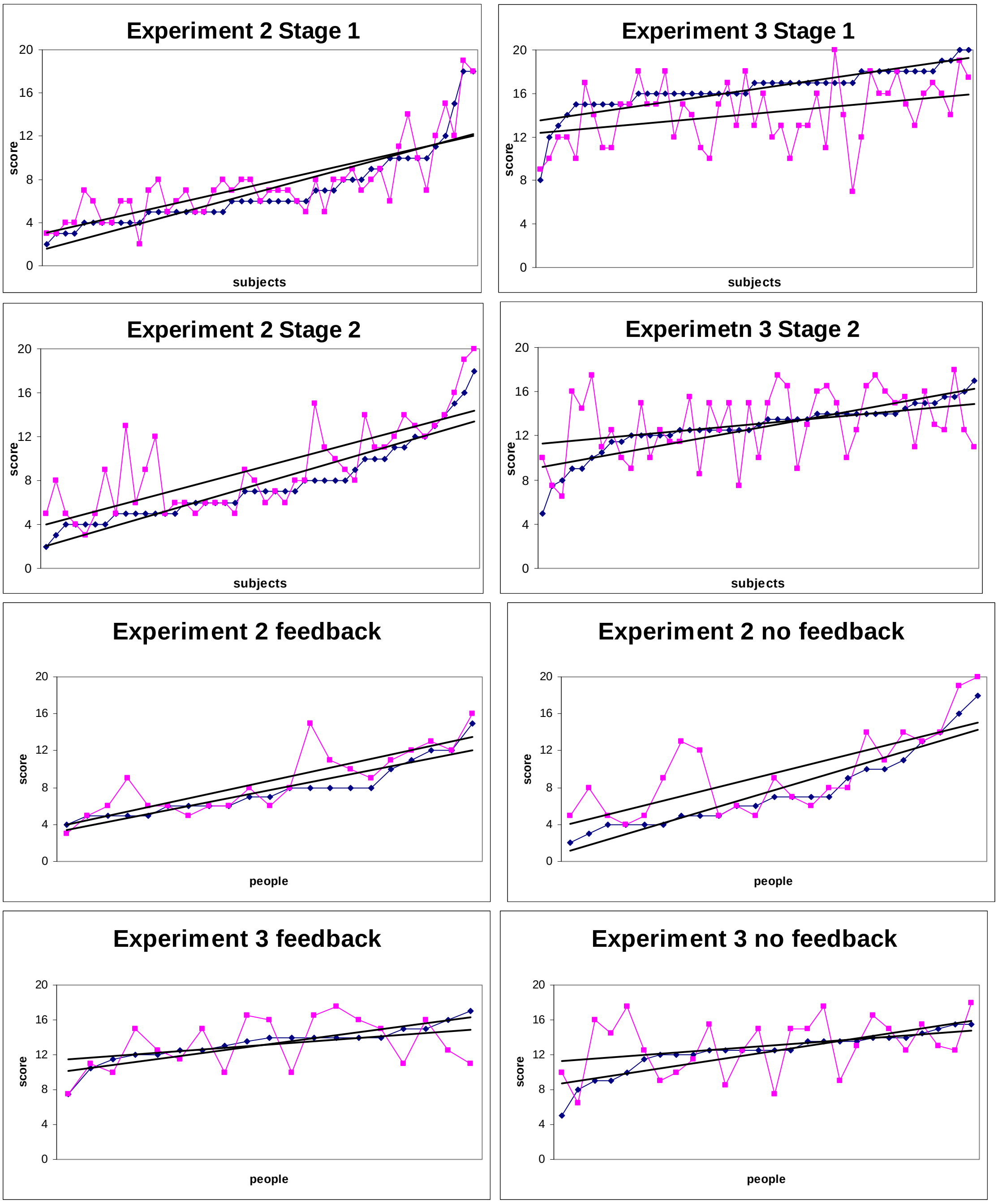


Graphs of actual percentiles (pink line) and estimated percentiles (blue line).



Appendix A2

Graphs of the real (blue line) and estimated (pink line) distribution of score (all data adjusted and ordered from the lowest real score to the highest real score); together with trend lines for each series.



Experiment 2 Stage 1

Experiment 3 Stage 1

Experiment 2 Stage 2

Experimetn 3 Stage 2

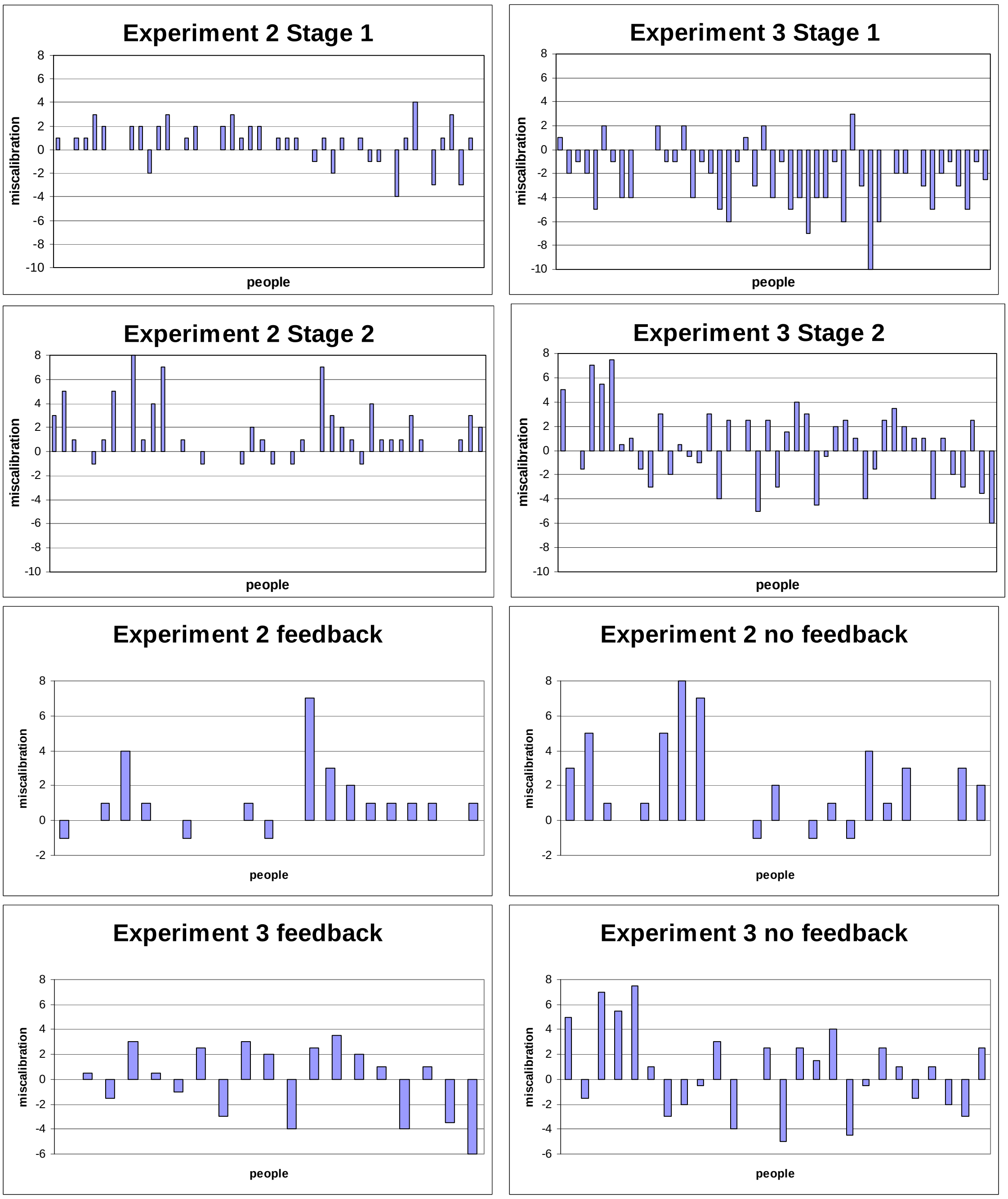
Experiment 2 feedback

Experiment 2 no feedback

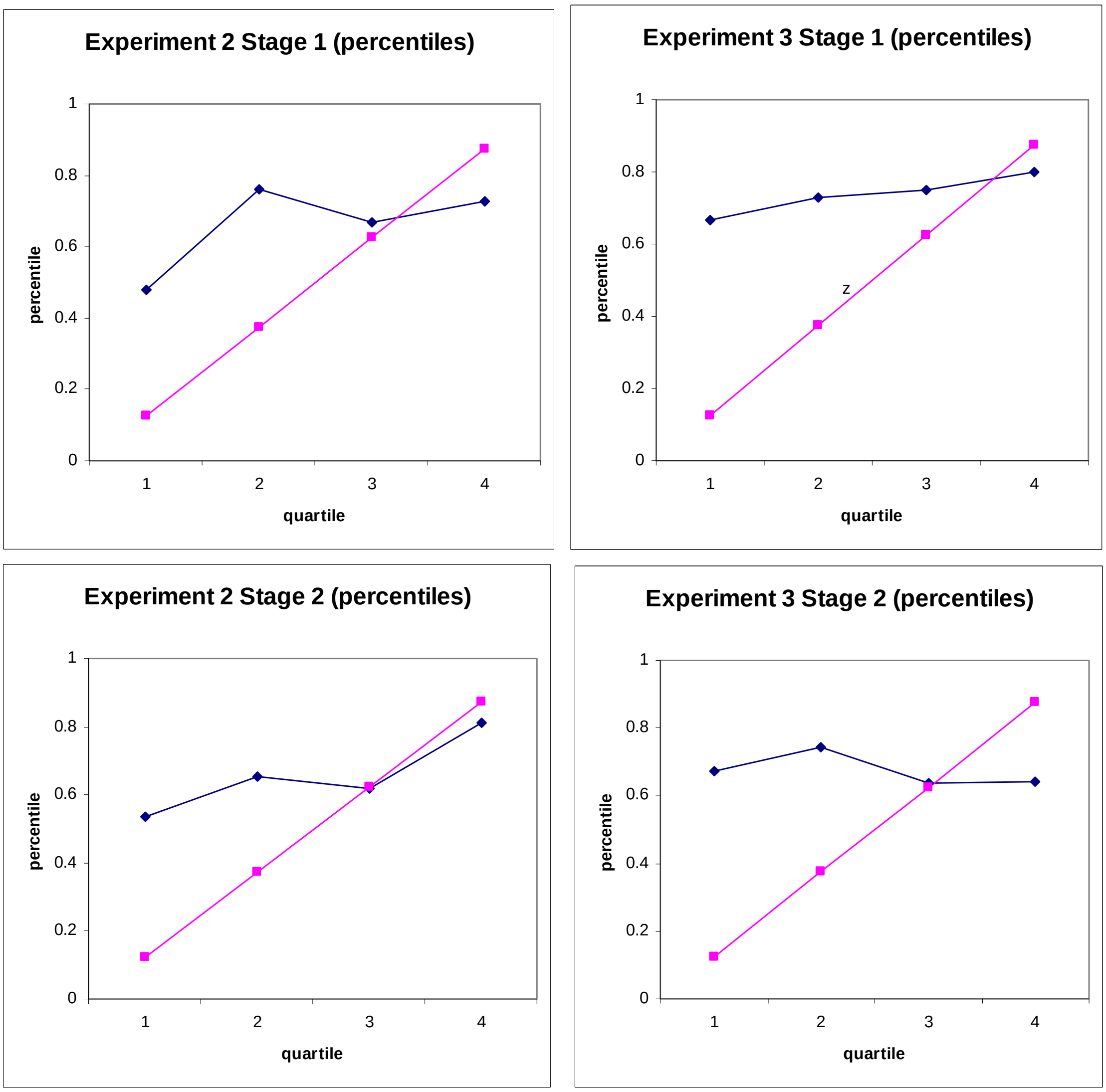
Experiment 3 feedback

Experiment 3 no feedback

Graphs of distributions of miscalibration of Own score (all data adjusted and ordered from the lowest real score to the highest real score).



Graphs of actual percentiles (pink line) and estimated percentiles (blue line).



Appendix A3

For each task we first computed the number of people whose calibration in Own score estimates improved/did not change/worsened in Stage 2 (compared to Stage 1)[64].

Table A3. *Number of people who improved//did not change/ worsened their calibration.*

Own score	Experiment 2			Experiment 3		
	pooled	feed	no feed	pooled	feed	no feed
Worse	11	7	4	16	6	10
Equally good	13	4	8	6	2	4
Better	16	8	9	18	11	7

These results reveal that there is more of those people who improved (or at least did not worsen) their calibration over time than those who worsened it. Unfortunately, the number of observations in both tasks is too low to reach statistical significance of the difference between better and worse performing people.

Appendix A4

We computed how many people, depending on the additional feedback, improved/did not change/worsened their calibration in Stage 2 (compared to Stage 1). Table A3 summarizes the results. These results suggest that additional feedback decreases miscalibration even more. Similarly as in case of Hypothesis 2a are these results only descriptive because due to the small number of observations we cannot show statistical significance of these differences.

[64] We computed these counts with absolute values of miscalibration, i.e. if someone's miscalibration in Stage 1 was 4 and in Stage 2 it was -3, then for this subject we identified improvement in calibration.

Individual researchers, as well as the on-line and printed versions of the CERGE-EI Working Papers (including their dissemination) were supported from the following institutional grants:

- Center of Advanced Political Economy Research [Centrum pro pokročilá politicko-ekonomická studia], No. LC542, (2005-2009),
- Economic Aspects of EU and EMU Entry [Ekonomické aspekty vstupu do Evropské unie a Evropské měnové unie], No. AVOZ70850503, (2005-2010);
- Economic Impact of European Integration on the Czech Republic [Ekonomické dopady evropské integrace na ČR], No. MSM0021620846, (2005-2011);

Specific research support and/or other grants the researchers/publications benefited from are acknowledged at the beginning of the Paper.

Published by
Charles University in Prague, Center for Economic Research and Graduate Education (CERGE)
and
Economics Institute ASCR, v. v. i. (EI)
CERGE-EI, Politických vězňů 7, 111 21 Prague 1, tel.: +420 224 005 153, Czech Republic.
Printed by CERGE-EI, Prague
Subscription: CERGE-EI homepage: http://www.cerge-ei.cz

Editors: Directors of CERGE and EI
Managing editors: Deputy Directors for Research of CERGE and EI

ISSN 1211-3298
ISBN 978-80-7343-173-0 (Univerzita Karlova. Centrum pro ekonomický výzkum
a doktorské studium)
ISBN 978-80-7344-162-3 (Národohospodářský ústav AV ČR, v. v. i.)

CERGE-EI
P.O.BOX 882
Politických vězňů 7
111 21 Praha 1
Czech Republic
http://www.cerge-ei.cz