CERGE
Center for Economic Research and Graduate Education
Charles University Prague

Overconfidence in Business, Economics, Finance, and Psychology: How Much of a Problem is It?

Marian Krajč

Dissertation

Prague, June 2009

Dissertation Committee:

Andreas Ortmann (CERGE-EI, Prague)

Dmitry Ryvkin (Florida State University, USA)

Ralph Hertwig (University of Basel, Switzerland)

Peter Katuščák (CERGE-EI, Prague)

Referees:

Erik Hoelzl (University of Vienna, Austria)

Gerlinde Fellner (Vienna University of Economics and Business, Austria)

Marian Krajč

Overconfidence in Business, Economics, Finance, and Psychology: How Much of a Problem is It?

Dissertation

Prague, June 2009

Acknowledgments

This dissertation could not have been written without my advisor, Andreas Ortmann, who encouraged and challenged me throughout the process of researching and writing this thesis. I am deeply indebted to him for all his help, stimulating suggestions, and always patient and careful reading and commenting on numerous drafts of my papers. He is in fact the person who inspired my interest in academic research and never accepting less than my best efforts.

I am also indebted to Ralph Hertwig, whose insightful comments and constructive criticisms at different stages of the research helped me focus my ideas.

My special thanks go to Dmitry Ryvkin for his advice and guidance in designing the experiments and in analyzing the data.

I would also like to thank Aldo Rustichini for his help during my mobility stay at the University of Minnesota.

Furthermore, I would like to thank Pavlo Blavatskyy, David Budescu, Levent Celik, Peter Katuščák, Bradley Ruffle, Louise Ryan, participants of the International Conference on Affect, Motivation and Decision Making (Ein Bogeg, The Dead Sea, Dec 12 – 15, 2006), and an anonymous referee for the Journal of Economic Psychology for valuable comments on the first chapter of this dissertation.

I am also grateful to Peter Katuščák, Petr Zemčík, and participants of the 8th GDN Regional Research Conference for their comments on the experimental chapter of this dissertation.

I appreciate the financial support of the CERGE-EI Foundation under a program of the Global Development Network that funded the experimental part of the research.

My friends and colleagues from CERGE-EI gave me much-needed support. I want to thank them for all their help, support, interest and valuable hints. In particular, I am obliged to Katarína Kálovcová, Natalia Shestakova, and Jan Mysliveček.

Most importantly, none of this would have been possible without the love and patience of my family, especially my wife Jana, whose patient love enabled me to complete this work. I am also grateful to my parents and my sister for their generous support over the years.

Contents

General introduction

Psychologists seem to have been the first to have identified overconfidence as a widespread phenomenon; however, economists' attention has most recently been drawn to it. Overconfidence, as well as its twin sibling underconfidence, has become one of the most investigated biases in the last two decades –experimentally, empirically, and theoretically. In general, the term overconfidence is used in situations when the observed estimations of values (of confidence or probabilities) are consistently greater than the real values. Overconfidence can be measured in a number of ways – e.g., comparing mean subjective probability or confidence in the estimates with the proportion of correct answers; comparing the estimated own ranking within a group with the real ranking; or comparing estimated confidence intervals with the true confidence intervals. Overconfidence is important in many areas of business, economics, and finance where overconfident behavior may cause substantial losses (e.g., entrepreneurs excessively entering markets, managers undertaking non-profitable projects, traders overestimating their abilities). The results of research in psychology as well as in business, economics, and finance are, however, ambiguous. Therefore, understanding this bias is essential in avoiding losses in various areas.

The dissertation consists of three chapters, all three of which investigate overconfidence from the view of experimental economics. The first chapter introduces a simple model that offers an alternative – non-biased – explanation to seemingly biased behavior which was experimentally identified in the literature. The second chapter reports three experiments in which the theoretical model is tested and the effects of various kinds of information on calibration are inspected. The third chapter reviews, categorizes, and evaluates the existing experimental literature on overconfidence, attaching significance to known issues from psychology and pointing out the main shortcomings of the reviewed literature.

The first chapter, which is a joint work with Andreas Ortmann and has since been published in the Journal of Economic Psychology, offers a theoretical model that suggests an alternative explanation to the so-called unskilled-and-unaware problem – that the unskilled are seemingly afflicted by a double curse because they also seem to be unaware of their (relative) lack of skills. Essentially, the unskilled overestimate their skills while the skilled underestimate (but less than the unskilled). The unskilled-and-unaware problem was experimentally identified about a decade ago and numerous authors have elaborated on this problem since – experimentally as well as theoretically. We propose that the alleged unskilled-and-unaware problem, rather than being one of biased judgments, is a signal extraction problem that differs for the skilled and the unskilled. Specifically, the unskilled face a tougher inference problem than the skilled. The model is based on two assumptions. First, we assume that skills are distributed according to a J-distribution, which can be regarded as an approximation of the very right tail of the IQ distribution. This assumption is reasonable given the typical subject pool used in the experimental studies of overconfidence – students from prominent US universities. Second, we assume an error term in own-ability perception, which is a common assumption in psychology models. Our simple model generates, by means of analytical computations, patterns similar to those identified in the previous experimental literature. We also discuss conditions under which the unskilled-and-unaware problem should disappear.

The second chapter reports the results of three experiments (one field, two laboratory) through which we tested the theoretical model and some informal extensions. Specifically, we examine the impact of general information and specific information (feedback) on the quality of absolute and relative self-assessment ("calibration") in various tasks (microeconomics exam, skill-oriented task, and general-knowledge oriented task). In our experiments, we used a specific subject pool – CERGE-EI preparatory semester students who are competitively selected students from their home universities around Central and Eastern Europe. This allowed us to observe the evolution of calibration in a group from its beginning, when the group members have very little information about other members. The simple model replicates some of the patterns identified in experiments; however, it does not always fit the data very well. Overconfidence behavior initially prevails in almost all settings. We find a strong positive effect of general information on calibration. We also show that calibration improves more when feedback is provided. Moreover, our results suggest that the absolute self-assessment improves more than the relative self-assessment and therefore conclude that the absolute self-assessment is more responsive to information. In our experiments we also show that it is the unskilled who improve their calibration the most. Based on the results, we conclude that information plays an important role in the absolute as well as the relative self-assessment and that the unskilled-and-unaware problem arises mostly due to the lack of information.

The third chapter reviews, categorizes, and evaluates experimental studies on overconfidence and self-assessment in business, economics, and finance. First, we review the main results of experimental research in psychology and highlight the main issues in psychology as well as current issues in economics. Then we create, using the Econlit and Web of Science databases and employing a clear and replicable selection rule, a non-opportunistic set of experimental studies from business, economics, and finance concerning overconfidence or self-assessment. We show that overconfidence studies in business, economics, and finance are much more heterogeneous than those in psychology. We identify nine paradigms (General-knowledge questions, Confidence intervals, Forecasting, Market-entry games, Auctions, Willingness to sell/buy, Information, Assessment of others, Self-awareness questions) and categorize the experimental studies from the business, economics, and finance literature according to those paradigms. For each paradigm we then review the corresponding studies and point out the shortcomings of each study, paying attention to issues identified in psychology as well as to issues already known in economics. Finally, we discuss the existing research for each paradigm and, based on the review, make suggestions for further research.

CHAPTER 1

Are the Unskilled Really That Unaware? An Alternative Explanation

Joint work with Andreas Ortmann

Abstract

In a series of articles and manuscripts (e.g., Kruger and Dunning, 1999; Dunning et al., 2003; Ehrlinger et al., 2005), Dunning, Kruger and their collaborators argued that the unskilled lack the metacognitive ability to realize their incompetence. We propose that the alleged unskilled-and-unaware problem – rather than being one of biased judgements – is a signal extraction problem that differs for the skilled and the unskilled. Specifically, the unskilled face a tougher inference problem than the skilled.

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

1.1 Introduction

In a recent series of articles and manuscripts, it has been argued that the unskilled are, in addition, unaware of their incompetence: "Not only do these people reach erroneous conclusions and make unfortunate choices, but their incompetence robs them of the metacognitive ability to realize it" (Kruger and Dunning, 1999[1], p. 1121). The unskilled, thus, are allegedly afflicted by a "double curse" (Dunning et al., 2003, p. 84). In a recent manuscript (Ehrlinger et al., 2005), the authors replicated their earlier results, addressing various published critiques of their work (e.g., Ackerman et al., 2002; Burson et al., 2006; Krueger and Mueller, 2002; Krueger and Funder, 2004). By and large, the authors seem to find support for their original contention.

The results by Dunning and Kruger warrant further investigation because they seem to be at odds with the results reported by Juslin et al. (2000), who – through a meta-analysis of 130 data sets – demonstrated that over- and underconfidence disappear for general-knowledge questions that employ representative stimuli. According to these authors, people are well-calibrated.[2] To the extent that general-knowledge questions have a skills component, the results of these two sets of authors therefore seem at odds.

We argue that the subject pools used by Dunning, Kruger, and their collaborators were not distributed uniformly, or at least symmetrically, but rather skewed toward the bottom. We show below with a simple model that the unskilled, rather than being more unaware than the skilled, face a tougher inference problem which, at least partially, explains their alleged lack of metacognitive ability.

The remainder of this chapter is organized as follows: In the following section we provide a brief review of the literature. We then sketch out the intuition guiding our model. In section 1.4, we provide a simple numerical example meant to illustrate our intuition. In section 1.5, we complicate that numerical example and show that our alternative explanation can explain the three stylized facts of Kruger and Dunning (1999) that constitute the unskilled-and-unaware problem. In section 1.6, we discuss the conditions under which we expect the unskilled-and-unaware problem to disappear, sketch out possible experimental tests to verify our conjecture, and report briefly on experimental follow-up work. Section 1.7 concludes.

1.2 A brief review of the relevant literature

[1] As of November 30, 2007, the Kruger & Dunning (1999) article has attracted more than 350 references on Google scholar.

[2] General-knowledge questions are different from the person-oriented tasks (asking, typically, about one's capabilities, in both an absolute and relative manner) that are used to generate the unskilled-and-unaware problem. Juslin et al. (2000, p. 394) note that "one sensible but as yet untested hypothesis" is that there are important differences between these two paradigms. We are not aware of any studies that could shed light on this issue but would find it surprising indeed if there were such differences. Chapter 2 of this dissertation seems to represent the first stab at testing this conjecture.

Juslin et al. (2000) demonstrated that over- and underconfidence, at least for an important experimental paradigm in psychology (general-knowledge questions), is an artifact: People tend to be calibrated reasonably well in situations that they have had a chance to experience repeatedly (Cosmides and Tooby, 1996) – general-knowledge questions almost by definition fulfilling that criterion – and that are fairly described by the stimuli materials. (Here "fair" describes whether the selected general-knowledge questions, say in city comparison tasks, reflect the ecological validity of the cues; see Hertwig and Ortmann, 2005).

In contrast, Kruger and Dunning (1999; see also Dunning et al., 2003) suggested that, across many intellectual and social domains, the subjects that perform the poorest (the unskilled) also lack the metacognition that would allow them to assess their deficiencies. The authors argue that this double curse of being unskilled and unaware induces the unskilled to dramatically overestimate their expertise, knowledge, skills, talents, etc.[3] The authors also suggested that the very skilled are somewhat, but less so, unaware of their skills. The ability-perception divergence is, however, much less prominent at the upper tail than at the lower tail; the authors attribute this phenomenon to "undue modesty" (Dunning et al., 2003, p. 85) which strikes us as an unpersuasive argument.

To sum, three stylized facts beg for explanation: first, the alleged overconfidence of the unskilled; second, the alleged underconfidence of the very skilled; and, third, the asymmetry of the alleged miscalibrations.

The original findings were built on student subjects' knowledge of grammar and logical reasoning, and their self-assessment of how humorous they are; these findings have since been replicated with different tasks (e.g., Dunning et al., 2003: classroom exams) and also different subject pools (Edwards et al., 2003: clerks evaluating their performance; Haun et al., 2000: medical lab technicians evaluating their on-the-job expertise; Parikh et al., 2001: medical students assessing their interview skills).

A number of authors have questioned the results by Kruger and Dunning (1999). Krueger and Mueller (2002) proposed that regression-to-the-mean (RTM) and the better-than-average (BTA) effect could, jointly, explain the three stylized facts constituting the unskilled-and-unaware problem. RTM, to recall, is a statistical artifact that occurs when variables such as ability and the perception of ability are imperfectly correlated, possibly because of measurement errors. The imperfect correlation between ability and perception of ability, and a regression slope of less than 1 (as observed in Kruger and Dunning, 1999), imply that not all of those in the lower quartile in ability are actually in the lower quartile in perception. Thus, the expected value of the lower quartile in perception will be greater than the average ability of the lower quartile in abilities. While RTM explains the first two stylized facts, it cannot explain the third (the asymmetry of the alleged miscalibrations). Krueger and Mueller (2002) therefore appeal, in addition, to the BTA

[3] Below we often use these terms interchangeably.

effect.[4] Using the BTA effect as part of an explanation seems, however, problematic as it is the explanandum rather than the explanans.[5]

Recently, Ehrlinger et al. (2005) addressed a number of criticisms and alternative explanations of the results in Kruger and Dunning (1999) by using real-world settings and financial and social incentives. In the first part of their manuscript (study 1 and 2), Ehrlinger et al. (2005) investigated whether the performers in the bottom quartiles overestimated their relative and absolute ability after they controlled for measurement errors in real-world situations (in-class exams, debate tournaments); the authors argued that the results of these two studies confirm the original findings of Kruger and Dunning (1999) and undermine the RTM and BTA explanation of Krueger and Mueller (2002).

In the second part of their manuscript, the authors conducted three studies (study 3, 4, and 5) to examine whether insufficient incentives for accuracy are the reason for the overconfidence of poor performers. The participants of the third study were recruited at a Trap and Skeet competition and were asked to assess their confidence in the answers they gave to questions asked on a test of gun knowledge and safety. Participants were promised an additional $5 payment for average confidence responses within 5% of their actual score on the test.[6] Since the number of years of experience with a firearm was reported to be 6-65, we deduce that the participants of the Trap and Skeet tournament were not students. Hence this study seemed to show that the Kruger and Dunning (1999) results generalize to populations other than undergraduate students (and also other tasks). In the fourth study, the authors conducted a similar experiment on a logical reasoning test with undergraduate students who were promised an additional payment of $100 if their estimate of their performance was within 5% of their actual score on the test.[7] In the fifth study, the authors investigated the impact of social incentives (making students accountable for their self-assessment) on the results of Kruger and Dunning (1999). The "accountable" group was told that their professor would interview them regarding the rationale of their answers on a multiple choice test. The results of these three studies

[4] The better-than-average effect is the alleged human tendency to overestimate one's achievements and capabilities relative to others. It is also informally known as the Lake Wobegon effect. Note that Garrison Keillor's idyllic town is also a fictional town, as is possibly the BTA effect.

[5] Krueger and Mueller (2002) also study the effects of various mediators and find that mediation has no, or lower, explanatory power than RTM and BTA together. Kruger and Dunning (2002) questioned the mediation results by Krueger and Mueller (2002), arguing they used unreliable tests and inappropriate measures of relevant mediating variables. They also point out that the results of Krueger and Mueller (2002) are true only if low or moderate levels of reliability are used and not in samples with highly reliable measures. Since this particular dispute is not of relevance to our argument, we do not pursue it in more detail here.

[6] However, the authors do not report the expected value of having a correct prediction, or what fraction of subjects actually earned the extra money. It is quite possible that the expected value was too low and that the financial incentives were simply insufficient (Hertwig and Ortmann, 2001; Rydval and Ortmann, 2004). In addition there is the basic problem that participants can strategically manipulate their answers to get more money. In the extreme, a student can deliberately answer all questions incorrectly (or not at all), make an estimate that he/she answered zero questions correctly, and cash in $10.

[7] Again, the authors did not report the expected value of a hit. Even though the reward for accuracy was large nominally, the expected value might still have been low and financial incentives therefore insufficient. The incentive compatibility argument of the preceding footnote applies here, too.

seemed to suggest that neither monetary nor social incentives affect the overestimation of performers in the bottom quartiles.

In the third part of their manuscript, Ehrlinger et al. (2005) investigated the sources of inaccuracy in performance estimates. Towards that end, the authors computed via regression analysis how people weigh their estimates of their raw score and estimates of the raw score of the average person when estimating how well they performed relative to others. They then did a "what if" exercise (i.e., conducted a counterfactual regression analysis), asking what each participant's percentile ranking would be if her or his raw score (or, the average person's score) estimate were replaced with the real value. The results of this analysis suggest that the participants are inaccurate due to mistaken beliefs about their own performance, rather than due to a misconception about the performance of others.

In a related article on the unskilled-and-unaware problem, Burson et al. (2006) also questioned the results of Kruger and Dunning (1999). They conducted three studies to examine their hypothesis that task difficulty matters and suggested a noise-plus-bias model.

For the first study, students were asked to answer quizzes with either 20 hard or 20 easy questions about the University of Chicago. The students were then asked to estimate the number of correct answers they gave, their percentile rank, and the difficulty of their quiz.[8] For the easier questions, the authors replicated the results of Kruger and Dunning (1999); for the harder questions they found that both the percentile estimates of low and high performers decreased so that the asymmetry (stylized fact three) disappeared – the low performers are as aware as the high performers in percentile estimates (and even more aware in their score estimates). For the second study, Burson et al. (2006) varied domains (5), question sets (10), and difficulty (2) for student subjects. The participants were asked to estimate their percentile ranking and the task difficulty for themselves and for the other participants. In the third study, the student subjects were asked to create as many 4-, 5-, and 6-letter words as possible from a 10-letter word. Again, the participants were asked to estimate their percentile ranking, their number of points, and the task difficulty. The results of the second and third study support the results of the first study by Burson et al. (2006): the skilled and unskilled are similarly unaware of how they perform relative to others and the top performers are better calibrated in the easier tasks and the bottom performers in the harder tasks.

Burson et al. (2006) proposed a noise-and-bias model that is, according to the authors, sufficient to explain the observed behavior. In their model, the noise is caused by task randomness (e.g. random variation, luck, distraction, fatigue) and diagnosticity of feedback (what kind of feedback people get during the experiments) – this part of the explanation resembles the RTM argument (and the earlier work by Erev et al., 1994); the bias in their model captures the task difficulty – this part of the explanation resembles the BTA argument for easier questions.

[8] The authors paid only a flat participation fee in all three studies. Therefore, the caveats about (the lack of) performance-based financial incentive applies (e.g., Hertwig and Ortmann, 2001).

Krueger and Mueller (2002) and Burson et al. (2006) agree that the reliability of measures plays an important role in the analysis. In addition, the task difficulty is the key in explaining the asymmetry in Burson et al. (2006). The relationship between task difficulty and task reliability is, however, unclear, as Krueger and Mueller (2002) report higher reliability in easier tasks and lower in harder tasks; Burson et al. (2006) report it vice versa.

To sum up, the experiments conducted by Kruger and Dunning (1999) and Ehrlinger et al. (2005) suggest that the unskilled overestimate their absolute abilities as well as their relative abilities. The authors argue that the unskilled are overconfident about their abilities. This overconfidence is explained as resulting from a lack of metacognitive ability to realize their deficiencies. The meta-analysis of Ehrlinger et al. (2005) suggests, furthermore, that the lack of insight into participants' skills is the reason for the excessively optimistic self-assessments of poor performers. It also seems widely acknowledged that noise is an indispensable part of any sensible explanation.

1.3 Our alternative explanation – the intuition

Drawing on empirical data relevant to the subject pools typically used, we propose an alternative explanation of the results of Kruger, Dunning, and their collaborators. Our key observation is that for almost all studies by Kruger, Dunning, and their collaborators traditional but hardly representative subjects – undergraduate (psychology) students from Cornell – were employed, i.e., a convenience sample rather than a representative sample of the population.

Students at Cornell, and similar schools such as the University of Chicago (e.g., Burson et al., 2006), are drawn from the outer upper tail of the normal distribution of student talent. Take the example of Cornell University: According to U.S. News & World Report the percentage of applicants admitted to Cornell University is 29%. Clearly, because of Cornell's reputation,[9] this is already a sample from a self-selected pool. It seems unlikely that high-school students from the lower half of the talent distribution would apply. A similar argument is likely to apply to University of Chicago students.

The talent distribution of the subject pool typically used in the experiments is therefore highly asymmetric and can be approximately captured by the J-distribution, which one can think of as a truncated (from below) normal distribution. This pattern can be seen in the IQ distribution, the most general measure of a person's cognitive abilities. The convex part of the upper outer tail of that distribution represents approximately the top 15% of the population. Since Cornell University accepts only 29% of the applicants and since its pool of applicants is self-selected, it seems likely that, save some legacy cases (i.e. students that – because their parents were alumni/ae – were admitted even though their

[9] According to the U.S. News & World Report web site, Cornell University is currently the 12th best university in the U.S.

qualifications are less than stellar), almost all Cornell students are located in the convex part of the upper outer tail of the normal distribution.

In addition, it is well-known from studies of grade inflation (Avery et al., 2003; Johnson, 2003; Lewis, 2006, chapter 5; Sabot and Wakeman-Linn, 1991) that grades at the undergraduate level have – with the notable exception of the natural sciences – become less and less differentiating over the years: more and more students are awarded top grades. For example, between 1965 and 2000 the number of A's awarded to Cornell students has more than doubled in percentage while the percentage of grades in the B, C, D and F ranges has consequently dropped (in 1965, 17.5 percent of grades were A's, while in 2000, 40 percent were A's).[10] This data strongly suggests that Cornell University experiences the same phenomenon of (differential) grade inflation that Harvard experiences (Lewis, 2006) and the schools discussed in Sabot and Wakeman-Linn (1991). The dramatic grade inflation documented for the humanities and social sciences devalues grades as meaningful signals specifically in cohorts of students that are newly constituted and typically draw on the top of high school classes. Inflated grades complicate the inference problem of student subjects that, quite likely, were students in their first year or in their first semester (Ortmann and Hertwig, 2002).

To model the lack of feedback resulting from grade inflation and, possibly, the fact that student subjects were students in their first year or in their first semester, we introduce an error term in own-ability perception. The presence of noise in people’s ability assessment has already been acknowledged in RTM explanations. Noise is likely to be correlated with familiarity 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. We assume the distribution of this error to be homogeneous across subjects although one could reasonably argue for heterogeneous errors, with larger errors being correlated positively with lack of skill. Ultimately, noise is a function indeed of task randomness as well as diagnosticity of feedback. Note that by assuming homogeneity of errors, we handicap ourselves.[11]

Below we show how abilities and perception of abilities drift apart as a function of both the noise (error) and the bounded asymmetric distribution of talent. The error in our analysis is initially assumed to be normally distributed; our argument seems robust to other symmetric specifications of the error term. In the next section, we will illustrate the basic idea with a very simple numerical example that matches the observed stylized facts reported by Kruger and Dunning (1999) and Ehrlinger et al. (2005).

The upshot of our model is that the students in the bottom quartile(s) face a tougher inference problem: it is more difficult for students in the bottom quartile(s) to estimate correctly their relative standing from the feedback that tightly clustered signals (grades) provide. By way of example, the A++ student does not face much of an inference problem, whereas an average B student does, especially if the task pertains to domains

[10] http://www.thehoya.com/news/041202/news7.cfm

[11] Heterogeneous errors would magnify overestimation of the unskilled.

where he or she is not likely to have had much previous feedback about their relative standing (e.g., tasks on grammar, logical reasoning, humor, and the like).

1.4 Our explanation – a simple numerical illustration

We propose the following simple formula for perceived ability (y):

$y = x + \varepsilon$,

where x is the true ability (distributed according to a J-distribution) and ε is the error term (normally distributed). Roughly, and for reasons of simplification (but inspired by the "binning" into four skills levels in Kruger and Dunning, 1999), in Table 1.1, we represent the J-distribution of true skills.

Table 1.1. *J-distribution of true skills (1 = the worst, 4 = the best).*

Skill rank	1	2	3	4
# of subjects	8	4	2	1

We represent the normal distribution of the error term in Table 1.2, where 0.3 stands for the probability of a correct self-assessment for each of the skill levels and 0.22 (0.13) for the probability of under- or overestimating their ability by one (two) level(s).

Table 1.2. *The normal distribution of the error term.*

Misclassification	-2	-1	0	+1	+2
Probability	0.13	0.22	0.3	0.22	0.13

Since we assume boundaries on skill categories (min=1, max=4), for subjects in the boundary categories it is possible to misclassify their ability only in one direction (e.g. someone with the lowest real skill level (1) can perceive herself as being of the same level (1), or of better level (2 or 3) only).[12] A similar logic applies to skill levels that are close to boundaries. We therefore truncate the probabilities for boundary and relevant interior categories by removing impossible events (skill levels outside the range 1-4) and by normalizing the remaining probabilities to sum up to one for each category. Thus, even though initially we assumed our error distribution to be normal (or at least symmetric), our restrictions on the support of perceived abilities obviously affect the error distribution for ability categories close to, or on, the boundaries (see Table 1.3).

Table 1.3. *The renormalized probabilities for the four skill levels after truncation.*

Misclassification	-2	-1	0	+1	+2
Probabilities for skill rank 1			0.46	0.34	0.2
Probabilities for skill rank 2		0.25	0.35	0.25	0.15
Probabilities for skill rank 3	0.15	0.25	0.35	0.25	
Probabilities for skill rank 4	0.2	0.34	0.46		

[12] This argument can be justified by the selective nature of the subject pool as well as the typical distribution of grades at most colleges and universities.

With this distribution of errors, the resulting perceived distribution of skills is described in Table 1.4 and in Figure 1.1.

Table 1.4. *Distribution of perceived skills (1 = the worst, 4 = the best).*[13]

Skill rank	1	2	3	4
Mass of subjects	4.98	4.82	3.64	1.56

Figure 1.1. *True and perceived distributions of skills – with 4 skill categories.*

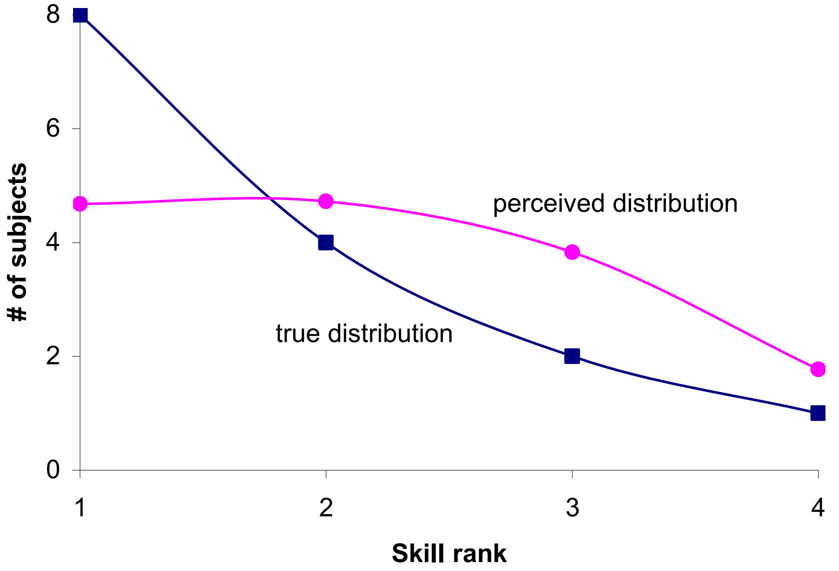


Figure 1.1 illustrates that as a consequence of the truncated error term, subjects perceive themselves as more skilled than they really are: the average perceived skill level increases from 2.88 for the true distribution to 3.26 for the perceived distribution. We see that this increase is driven mostly by subjects in the bottom quartiles – they perceive themselves as having better skills. This is the essence of what Kruger and Dunning (1999) and Ehrlinger et al. (2005) observed: about half of the subjects of skill rank 1 will move up (leaving four subjects – or .46 of the eight subjects in that category – in the bottom skill rank in perception), while one quarter of the subjects of skill rank 2 (or .25 of the four subjects in that category) will move down to the bottom. Plus a fraction of a subject of skill rank 3 (or .15 of 2 subjects) will also move down to the bottom.

Note that our simple example shows that the unskilled overestimate their abilities. Based on this bias in perceived abilities, people assess their relative ranking in the group. Therefore, we predict that a similar pattern can be found in percentile rankings.

While our simple example exhibits a similar pattern in the bottom quartiles as the results of Kruger and Dunning (1999), it does not capture behavior in the top quartile. As we

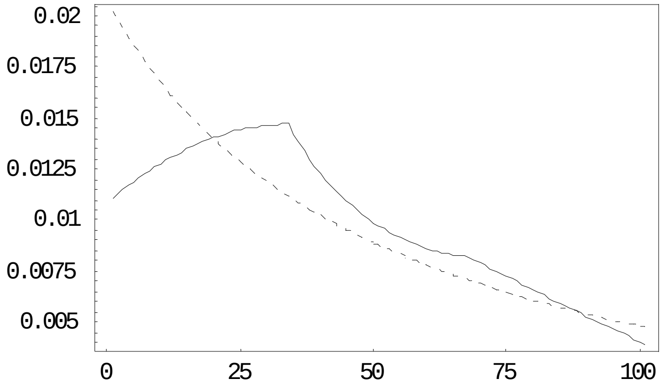
[13] The mass of subjects in each skill level was computed according to the simple formula for perceived ability (y) using the probabilities of truncated errors. For example, for perceived skill level 2, 4.82 = 0.35*4 + 0.34*8 + 0.25*2 + 0.2*1=(subjects with true skills 2 and perceived skills 2)+(true 1 and perceived 2)+(true 3 and perceived 2)+(true 4 and perceived 2).

will see presently, this is an artifact resulting from the low number of skill levels considered in this simple numerical illustration.

1.5 Our explanation – the example generalized to more skill levels

In this section we generalize the simple example to more levels of abilities. To generate a J-shaped distribution of real abilities we use one of the J-shaped distributions – the Chi-square distribution.[14] The error term is, as in our illustration, generated by the normal distribution which is, however, truncated at (or close to) the edges of the abilities range (where close is defined by the width of the error). Perceived ability is then computed as real ability plus (renormalized) error in one's self-assessment.

Figure 1.2. *True and perceived distributions of skills – with 101 skill categories.*



On the x-axis in Figure 1.2, the range of abilities runs from 1 (the lowest) to 101 (the highest). The y-axis corresponds to the number of subjects (where total mass of subjects is 1). The dashed line is the probability density function of the real abilities and the solid line is the probability density function of the perceived abilities constructed as described. We again observe a significant shift of the mass of subjects with lower abilities (dashed line) towards higher perceived abilities (solid line). We also, however, observe a shift of the mass of subjects with very high abilities towards lower perceived abilities. This captures the second stylized fact constituting the unskilled-and-unaware problem. The third stylized fact also emerges as the misperception of the unskilled is much larger than the misperception of the very skilled.

The distribution of perceived abilities is, of course, dependent on the choice of the key parameters: standard deviation and truncation of errors (i.e. for how many ability categories we allow people to make errors). We have conducted various robustness tests

[14] We are using the Chi-square distribution as a proxy to the J-distribution because the Chi-square distribution is, unlike the J-distribution, parameterized. Another possibility is to use the Pareto distribution, which also has the desired shape. The results obtained using the Pareto distribution are very similar to the results we get using the Chi-square distribution.

and find that our basic result is robust to variations in the standard deviations and truncation of errors.[15]

Our model shows what is intuitively clear: the less error we assume (i.e., the less complicated is the inference problem that subjects are assumed to face), the more likely people are to make accurate predictions and the less likely they are to fall victim to the unskilled-and-unaware problem. These results are consistent with the results of the meta-analysis in Ehrlinger et al. (2005), where the authors show that correcting the errors in own perceived ability helps people to assess their percentile ranking more accurately.

Prompted by a referee, we have also conducted computational robustness tests on our twin assumptions of a J-distribution of talent and the boundedness of the talent distribution. Specifically, we asked what would happen if the probability density function were not decreasing for the whole range. Even more specifically, what would happen if the distribution of abilities contained a tapering drop-off at the left bound (reflecting possibly legacy cases) rather than the sharp drop-off that we assumed so far for ease of exposition? We can show that there are parameterizations that allow us to reproduce the first of our stylized facts under these conditions. These parameterizations thus justify the more general assumption that the number of subjects with higher abilities is smaller than those with lower abilities. Thus, our assumption of a J-distribution of talent is less stringent than it might seem. Note that this result also speaks to the twin assumption of boundedness of the talent distribution because allowing the J-distribution to drop off taperingly rather than sharply at the left bound essentially moves the left bound further to the left.

Not surprisingly, both the slope of the distribution and the curvature of the distribution play a role, too: we can show that our original assumption of a J-distribution can be weakened considerably to generate stylized facts one and three. Of course, there are also parameterizations of the distribution where these stylized facts cannot be reproduced any longer. However, such parameterizations are rather extreme (e.g., for example, our model does not generate underestimation of abilities of the skilled when the number of skilled becomes too small relative to the number of unskilled, an effect that might get reinforced if, in addition, we assume the distribution to be convex).[16]

The referee also questioned if our model works in the case of seemingly unbounded measures (e.g., when the range of score is 0-100 and all people score between 40 and 60). One answer is that people have an intuitive understanding of subjective (hypothetical) bounds for all tasks and domains. Given the task, people first judge the task difficulty and the quality of the group members and thus generate expectations that no one would score below some level (lower bound) and no one would score above some level (upper bound). We have tested this assumption experimentally (see Chapter 2 of this dissertation) and the results suggest indeed that people have well-calibrated ideas about the minimum and maximum score achieved in the group. Therefore, our model is applicable to seemingly unbounded cases.

[15] These computations are available from the corresponding author.

[16] The details of this analysis are available on request from the corresponding author.

1.6 Discussion

The following table summarizes subject pools, financial incentives, and real-world stimuli in the two sets of studies that have motivated our inquiry (Kruger and Dunning, 1999; Ehrlinger et al., 2005).

Table 1.5. *Summary of Kruger and Dunning (1999) and Ehrlinger et al. (2005).*

Study	Real-world stimuli?	Financial incentives?	Subject pool
K&D(1999)[17] – Study 1	(humor)	No (extra credit)	CU undergraduates*
K&D(1999) – Study 2	(logical reasoning)	No (extra credit)	CU undergraduates*
K&D(1999) – Study 3	(English grammar)	Yes ($5 or extra credit)	CU undergraduates
K&D(1999) – Study 4	(logical reasoning)	No (extra credit)	CU undergraduates
E(2005)[18] – Study 1	(performance on in-class exam)	No (extra credit)	CU undergraduates*
E(2005) – Study 2	(debate tournament)	No	Students participating in a debate tournament
E(2005) – Study 3	(Trap and Skeet)	Yes	?
E(2005) – Study 4	(logical reasoning)	Yes	CU undergraduates*
E(2005) – Study 5	(logical reasoning)	No (extra credit)	CU undergraduates*

* all psychology students

We have argued that the asymmetric distribution of talent and the boundedness of the talent distribution (especially at the lower end) that we are likely to find in frequently used samples such as those drawn from elite colleges and universities like Cornell or Chicago (and also of places like CERGE-EI, see Chapter 2 of this dissertation), are key determinants of the experimental findings that constitute the unskilled-and-unaware problem. Indeed, with one exception,[19] all studies reported in Kruger and Dunning (1999) and Ehrlinger et al. (2005) featured student subjects.

What about participants like those in the Trap and Skeet competition or other subject pools used in related studies? An explanation in those cases is more difficult as both the asymmetric distribution of talent and the boundedness of the talent distribution are more difficult to justify for these samples. We conjecture, however, that the twin assumptions of an asymmetric distribution of talent and the boundedness of the support of the talent distribution (especially at the low end) is likely to be encountered frequently (for

[17] Kruger and Dunning (1999)

[18] Ehrlinger et al. (2005)

[19] Ehrlinger et al. (2005) recruited participants at a Trap and Skeet competition in some nearby club; these participants had 6-65 years of experience with firearms (mean=34.5), 96% owned at least 1 firearm and 89% had taken a course in firearm safety. We have no information about the actual distribution of skills in this study and we do not know how much the subjects in this competition could reasonably make inferences about their absolute and relative skills on a test that seems not have been related to the task they were in the process of performing.

example, for medical students that tend to be a highly selected group but also for participants in Trap and Skeet competitions, or other sports events.) Recall that we have weakened these assumptions in the previous section. If indeed these weakened conditions are fulfilled, then we expect the three stylized facts to emerge in all those situations where people are not given feedback about where they stand in the grand scheme of things. This is most likely to happen in situations where groups or cohorts are newly constituted from selected applicants that do not (yet) have, and do not have repeatedly, precise signals about their performance and their relative position in their group or cohort.

Our explanation suggests that the alleged unskilled-and-unaware problem may be an experimental artifact but for different reasons than conjectured in the literature so far. Unskilled students do not necessarily[20] lack metacognitive ability any more than more skilled students; they simply face a tougher inference or signal extraction problem. This inference problem can be mitigated through information about one's own position in the grand scheme of things. Where information/feedback about one's own position – e.g., through repeated and precise feedback such as ELO numbers or golf handicaps – is available, the inference problem gets attenuated. In our view it is exactly the lack of information/feedback about one's own position that makes it difficult for student subjects to make the appropriate inferences about where they stand relative to others. We suspect that these arguments also apply to those other populations and tasks that have been used in the literature: medical students assessing their interview skills; medical lab technicians evaluating their on-the-job expertise; clerks evaluating their performance.

Our suggestion fits well in an established body of literature that suggests that people, when allowed and able to learn, will do so (e.g., Koehler, 1996; see also Kruger and Funder, 2004). Take a cause celebre to both economists and psychologists: Chu and Chu (1990) and Cox and Grether (1996) have shown experimentally that it takes only a couple of rounds of repetition coupled with financial incentives and feedback for the preference reversal phenomenon to be driven out. Needless to say, the precision (diagnosticity) of the feedback, as well as the financial and social incentives, will affect subjects' ability to learn.

Our explanation suggests two (experimental) tests. First, an experimenter might want to ask early in the semester and late in the semester students in a class – especially a newly constituted class – about their absolute performance and relative standing: It seems highly likely that subjects in such a situation will quickly learn where they stand even when subjects start out with homegrown, and misleading, priors because they come from varied backgrounds and may have been selected because of their relative quality in the home setting. How fast miscalibrations will disappear is likely to be moderated by the quantity and quality of feedback (e.g., the number of midterms, the announcements of the distribution of grades of the midterm(s), but also the initial distribution of talent).

We report (see Chapter 2 of this dissertation) the results of exactly such an experiment: it employs applicants for a graduate school program that participated in a two-month

[20] They might but this is a topic of current research.

preparatory semester. Using the performance and predictions of performance on a midterm and a final as a natural field experiment and embedding in this field experiment a laboratory experiment in which we study, among other things, the speed of adjustment in misperceptions resulting from general and specific information, we find that even though students' calibration in exam score and percentile ranking predictions is poor at the beginning of the semester (not surprisingly since the distribution of scores on skill-related tasks is indeed very similar to a J-distribution that drops off taperingly at the left bound), it improves considerably throughout the semester, especially after the midterm results are revealed.[21] We also identify that the speed of adjustment in misperceptions is in both – own score and percentile rankings – higher if specific information (in the form of full feedback about absolute performance, relative standing, and average group score from the previous stage) is provided. Finally, an interesting result is the positive effect of representative stimuli on calibration: in Chapter 2 of this dissertation we show that miscalibration (absolute as well as relative) is stronger in the field experiments (exam predictions) than in the laboratory experiments where representative stimuli were implemented.

Second, we conjecture - because of the time they have to ascertain their absolute and relative performance as well as the differential diagnosticity of feedback that the well-documented differential grade inflation at US colleges and universities implies – that we will see a significant difference in the quality of self-assessments of first-semester/year humanities or social science students on the one hand and last-semester/year natural sciences majors even at elite schools like Cornell and Chicago on the other hand.

1.7 Concluding remarks

Dunning, Kruger, and their collaborators have proposed that the unskilled suffer from the "double curse" of being unskilled and being afflicted by a lack of metacognitive ability to realize their deficiencies. We have provided an alternative explanation: the unskilled face a much tougher inference or signal extraction problem. In other words, we suggest that flawed self-assessments do not necessarily result from biased judgements but can be explained as a signal extraction problem that differs for the skilled and the unskilled.

Our results seem of importance to the unskilled and unaware as well as to those who try to understand, and remedy, the situation. Specifically, as also demonstrated by the results on skills in Chapter 2 of this dissertation, it seems to take little feedback to significantly reduce over- and underconfidence relatively quickly if the information is precise and meaningful. By and large, our results seem to lend support to the results reported by Juslin et al. (2000), who demonstrated that over- and underconfidence disappear for general-knowledge questions that employ representative stimuli.

[21] We report large information effect size (Cohen's d) between miscalibration from midterm and final exam score predictions and medium effect size in the case of percentile ranking predictions.

Our results, thus, make a fundamental methodological point: little can be said about (the lack of) metacognitive ability if one does not control for the distribution of real abilities, task randomness, the diagnosticity of feedback, and financial and other incentives.

References

Ackerman, L.P., Beier, E.M., and Bowen, R.K., 2002. What We Really Know about Our Abilities and Our Knowledge. *Personality and Individual Differences, 33*, 587-605.

Avery, Ch., Fairbanks, A., and Zeckhauser, R., 2003. *The Early Admissions Game: Joining the Elite.* Harvard University Press.

Burson, A.K., Larrick, P.R., and 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.

Chu, Y., Chu, R., 1990. The Subsidence of Preference Reversals in Simplified and Marketlike Experimental Settings: A Note. *The American Economic Review, 80*, 902-911.

Cosmides, L., Tooby, J., 1996. Are Humans Good Intuitive Statisticians After All? Rethinking some Conclusions from the Literature on Judgment under Uncertainty. *Cognition,* 58, 1-73.

Cox, C.J., Grether, M.D., 1996. The Preference Reversal Phenomenon: Response Mode, Markets and Incentives. *Economic Theory, 7*, 381-405.

Dunning, D., Johnson, K., Ehrlinger, J., and Kruger, J., 2003. Why People Fail to Recognize Their Own Incompetence. *Current Directions in Psychological Science, 12*, 83-87.

Edwards, K.R., Kellner, R.K., Sistrom, L.Ch., and Magyaria, J.E., 2003. Medical Student Self-Assessment of Performance on an Obstetrics and Gynecology Clerkship. *American Journal of Obstetrics and Gynecology, 188,* 1078-1082.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., and Kruger, J., 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.

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

Haun, D.E., Zeringue, A., Leach, A., and Foley, A., 2000. Assessing the Competence of Specimen-Processing Personnel. *Laboratory Medicine, 31*, 633–637.

Hertwig, R., Ortmann, A., 2001. Experimental Practices in Economics: A Challenge for Psychologists? *Behavioral and Brain Sciences, 24*, 383-403.

Hertwig, R., Ortmann, A., 2005. The Cognitive Illusions Controversy: A Methodological Debate in Disguise That Matters To Economists. In R. Zwick and A. Rapoport (eds.), *Experimental Business Research* (pp. 361-378). Boston, MA: Kluwer.

Johnson, E.V., 2003. *Grade Inflation: A Crisis in College Education.* New York: Springer.

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

Koehler, J.J., 1996. The base rate fallacy reconsidered: Descriptive, Normative, and Methodological Changes. *Behavioral and Brain Sciences, 19*, 1-53.

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.

Kruger, J., Dunning, D., 2002. Unskilled and Unaware – but Why? A Reply to Krueger and Mueller. *Journal of Personality and Social Psychology, 82*, 189-192.

Krueger, I.J., Funder, C.D., 2004. Towards a Balanced Social Psychology: Causes, Consequences and Cures for the Problem-Seeking Approach to Social Behavior and Cognition. *Behavioral and Brain Sciences, 27,* 313-376.

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.

Lewis, R.H., 2006. *Excellence Without a Soul: How a Great University Forgot Education*. PublicAffairs.

Ortmann, A., Hertwig, R., 2002. The Costs of Deception: Evidence from Psychology. *Experimental Economics, 5*, 111-131.

Parikh, A., McReelis, K., and Hodges, B., 2001. Student Feedback in Problem Based Learning: a Survey of 103 Final Year Students across Five Ontario Medical Schools. *Medical Education, 35*, 632-636.

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

Sabot, R., Wakeman-Linn, J., 1991. Grade Inflation and Course Choice. *The Journal of Economic Perspectives, 5*, 159-170.

CHAPTER 2

Are the Unskilled Really That Unaware? Understanding Seemingly Biased Self-Assessments[1]

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.

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

[1] This chapter was published as WP #373 in the CERGE-EI Working Paper Series.

2.1 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 (see Chapter 1 of this dissertation) that explains the unskilled-and-unaware problem with asymmetric distribution of skills and errors in judgement. 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 in Chapter 1 of this dissertation 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 2.2, we review the literature concerned with the unskilled-and-unaware problem and related issues. In Section 2.3, we motivate, detail, and enumerate our research objectives and research strategy. Section 2.4 describes the design and implementation of the experiments. In Section 2.5, we present the results. We discuss our results and conclude in Section 2.6.

2.2 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[2] 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.

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.

We offered an alternative explanation of the unskilled-and-unaware problem in Chapter 1 of this dissertation: We 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 model is based on two assumptions. First, we claim that the distribution of students' skills[3] is bounded and

[2] Percentile is not a 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.

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

that skills have a J-distribution[4]. Second, we assume that the self-assessment process involves unsystematic noise[5]. Employing these two assumptions, we 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. We also demonstrated that the first assumption can to some extent be weakened, while the qualitative results remain the same

In Chapter 1 of this dissertation we also discussed the conditions under which they expect the unskilled-and-unaware problem to disappear. We 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).

2.3 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), excessive entry of the unskilled and 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. (2005) 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

[4] 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.

[5] 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" (Chapter 1 of this dissertation, p. 12).

explanations for the contradictory results of Camerer and Lovallo (1999) and Elston et al. (2005). 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. (2005) 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 in Chapter 1 of this dissertation 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 in Chapter 1 of this dissertation. We test the assumptions of the model (J-distributed score and error in judgement) as well as its performance in generating ability perception.

Hypothesis 1: *The **model** from Chapter 1 of this dissertation 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 judgements and decisions (e.g., Duffy and Hopkins, 2005; Engelmann and Strobel, 2000). We 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.[6] 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: *There is 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

[6] 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).

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. (2000)[7] 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.[8]

[7] 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.

[8] 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.

2.4 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.

Table 2.1. *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.

2.4.1 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).

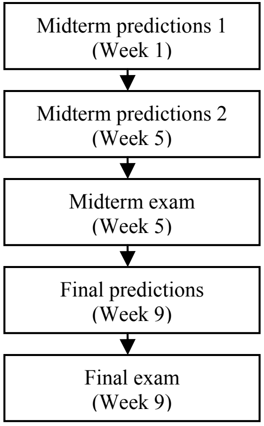
2.4.1.1 Design

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[9] and the average score as well as the

[9] 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

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).[10] Figure 2.1 illustrates the basic structure of Experiment 1.

Figure 2.1. *The structure of Experiment 1.*



2.4.1.2 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.[11] 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.[12] For each question, the participant with the best prediction was rewarded with 500 CZK.[13] All participants were told that their predictions would not affect their grades and that no one but the researchers would see the data.

We asked our subjects to predict their performance[14] ("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[15] ("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

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.

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

[11] 2 students came to the midterm exam after the questionnaire with predictions had been collected.

[12] 1 student did not hand in the exam sheet and 1 student came late.

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

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

[15] We will call this measure "Percentile" throughout the paper.

of the theoretical model from Chapter 1 of this dissertation, 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 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).

2.4.2 Experiments 2 and 3

With these experiments we addressed hypotheses 1 through 4.

2.4.2.1 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).

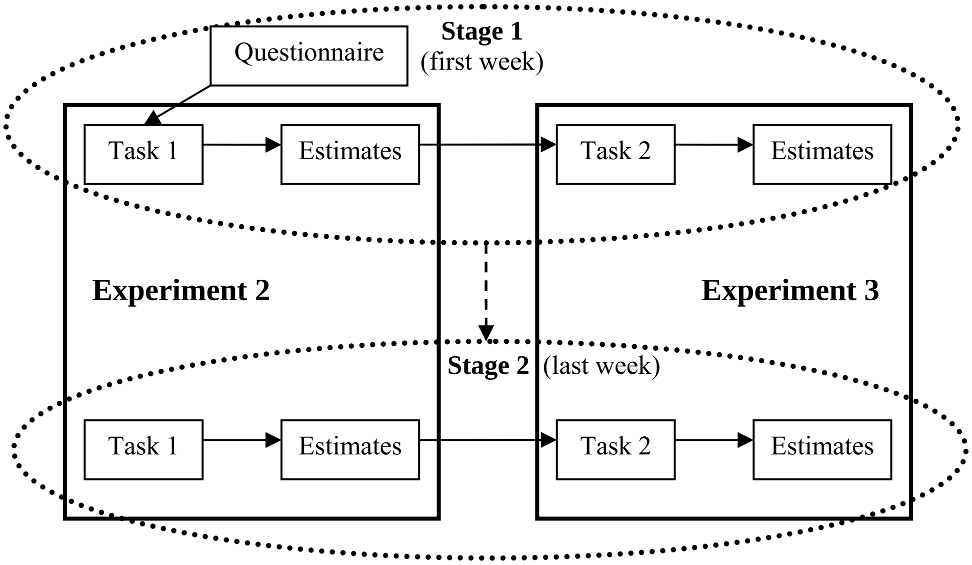
Task, Experiment 3: Participants had to complete, within a 2-minute time limit, a quiz containing 20 two-alternative general-knowledge questions, a research strategy widely investigated in psychology.[16] In Stage 1, we asked for a comparison of the population of

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

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.[17] 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 2.2.[18]

Figure 2.2. *The structure of Experiment 2 and Experiment 3.*



2.4.2.2 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.

[17] 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.

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

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.[19]

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.[20] 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)[21]. 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)[22] 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. 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 (see Chapter 1 of this dissertation) generates similar patterns as the experiment (hypothesis 1: "*model*"). We

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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; Rydval and Ortmann 2004; 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.

2.5 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 A (Experiment 1) and B (Experiments 2 and 3).

As we are primarily interested in miscalibration, we will mostly refer to miscalibration. All statistics in Tables 2.2a, 2.2b, and 2.2c (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)[23] and vice versa for percentage of better

[23] 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.

performing people (Percentile)[24]. Thus, a positive number means (on average) overestimation of the particular variable.[25] Note that while Own and Average scores are measured in scores, Percentile is measured in percentage.

Table 2.2a. *Experiment 1.*[26] *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

Table 2.2b. *Experiments 2 and 3, Stage 1.*[27] *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 2.2c. *Experiments 2 and 3, Stage 2. 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

[24] 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).

[25] 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.

[26] 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).

[27] In Stages 1 and 2, 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.

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

From Tables 2.2(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.

2.5.1 Hypothesis 1 - model

The model from Chapter 1 of this dissertation generates patterns of miscalibration similar to the predictions/estimates of Own score observed in the experiment.

2.5.1.1 Experiment 1

Recall that in Chapter 1 of this dissertation we 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, we 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 in Chapter 1 of this dissertation[30] on real distribution, compute the “simulated perceived score distribution” for every possible error width.

[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.

[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.

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

Table 2.3. *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 2.3 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 judgement 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.

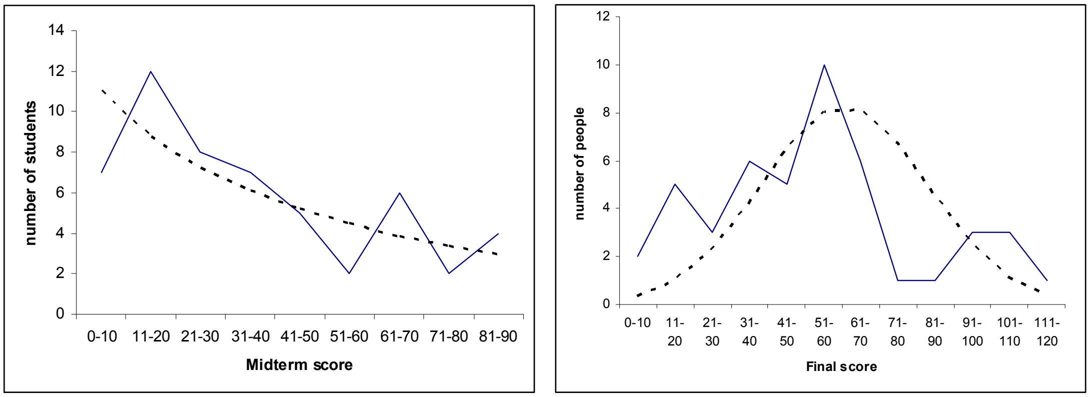
[30] To recall, the model assumes real distribution of score x and an error in perception ε. (Simulated) perceived distribution y is then formed as: $y = x + \varepsilon$. We did this with various widths of error.

[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).

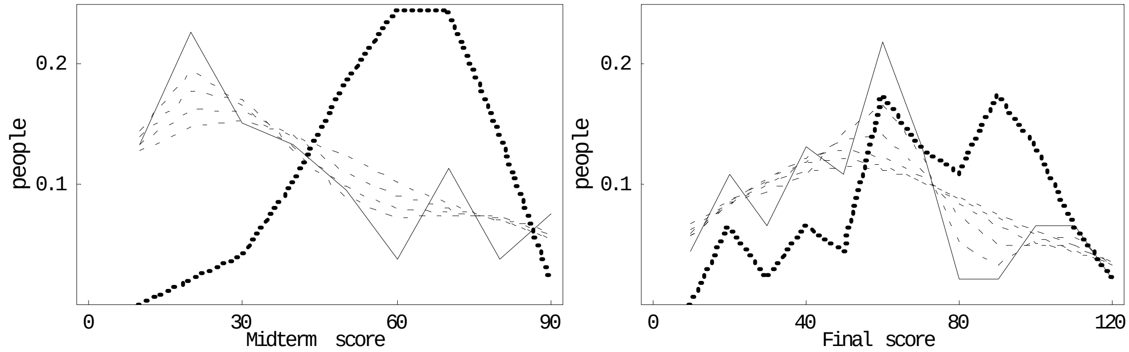
[33] For the same reason, we grouped abilities into intervals of 10 categories, thus we got 9 categories (this yields on average over 5 people in 1 new ability category).

Figure 2.3. *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).*



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 we conjectured in Chapter 1 of this dissertation 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 2.4. *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.

2.5.1.2 Experiments 2 and 3

Table 2.4 summarizes the results of simulations informed by the model proposed in Chapter 1 of this dissertation (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 2.4. *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

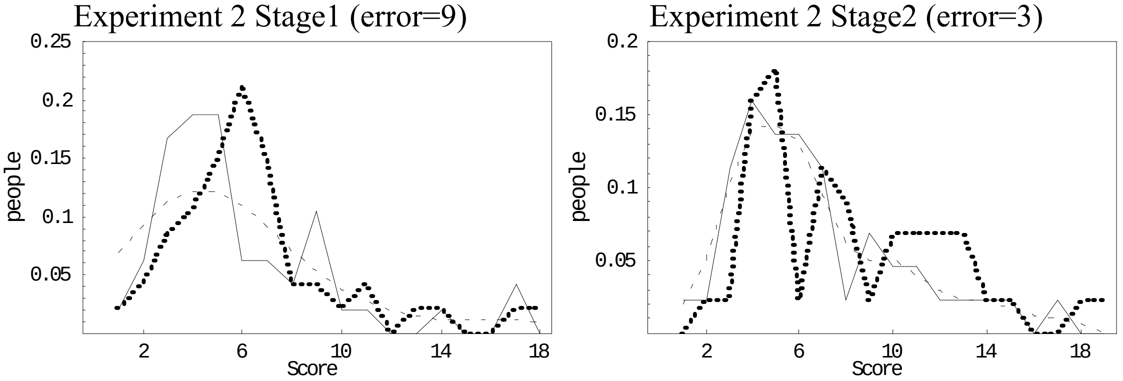
* - The number of ability categories over which people were distributed
** - Width of the error where the model fit is the best
*** - Minimum sum of absolute differences (best fit)

Table 2.4 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 2.5 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.

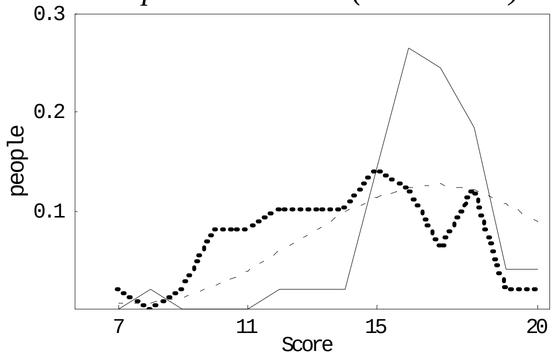
Figure 2.5. *The distribution of people over real score (solid line), experimental perceived score (dotted line), and simulated perceived score (dashed line) in Experiment 2 Stage 1 and Stage 2.*


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 2.6. *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).

2.5.2a Hypothesis 2a - general information

General information decreases miscalibration.

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

2.5.2a.1 Experiment 1

Descriptive results

Own. Table 2.5 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.

Table 2.5. *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].

[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).

[39] These numbers can also be found in Table 2.2a.

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 A.

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 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.

Own. 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 2.6. *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**

[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.

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.

Table 2.7. *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 2.8. *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.

Table 2.9. *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

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

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 the observed improvement of calibration over time statistically significant.

2.5.2a.2 Experiments 2 and 3

Descriptive results

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

Table 2.10. *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 2.10 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

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

[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.

3 pretty well calibrated in Own score estimates already in Stage 1 (see the graphs in Appendix B).[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.

Statistical results

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 2.11. *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.

[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 C.

[49] These numbers can also be found in Tables 2.2b and 2.2c.

[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).

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

Table 2.12. *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.

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.

Percentile. 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 2.13. *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 2.14. *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.

2.5.2b Hypothesis 2b – feedback

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

2.5.2b.1 Experiment 1

Not applicable.

2.5.2b.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.

Descriptive results

Own. Table 2.15 displays the average overconfidence and standard deviations of miscalibration of subjects with and without feedback in Stage 2.[52]

Table 2.15. *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 2.15 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]

Percentile. Table 2.16 displays the average overconfidence and standard of miscalibration of subjects with and without feedback in Stage 2.[55]

[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 D.

[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.

Table 2.16. *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 2.16 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 descriptive results suggest that, on average, subjects with feedback in Stage 2 outperform (in calibration) all subjects in Stage 1 in both tasks.

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.

Table 2.17. *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 turned out to be insignificant.[57] 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.

[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.

[57] The reason can be the small number of observations.

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 2.18. *Cohen's d.*

	Experiment 2	Experiment 3
p-value	**0.0484**	**0.0848**
Cohen's d	0.54 (0.26)[58]	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.

2.5.3 Hypothesis 3 – Own score vs. Percentile

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

2.5.3.1 Experiment 1

Descriptive results

We first expressed miscalibration of Own score predictions in percentage[59] 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 visually inspecting the basic statistics we could say that, in time, miscalibration is declining in both estimates and is also stabilizing.

Table 2.19. *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

Statistical results

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

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

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

Table 2.20. *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.

2.5.3.2 Experiments 2 and 3

Descriptive results

Visual inspection of the adjusted results in Table 2.21 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 2.21. *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

Statistical results

We used the same test (WSR test) as in Experiment 1 for testing for the difference in Own score and Percentile estimates.[60]

Table 2.22. *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

[60] 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.

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.

2.5.4 Hypothesis 4 – skills vs. general knowledge

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

2.5.4.1 Experiment 1

Not applicable.

2.5.4.2 Experiments 2 and 3

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, based on the basic statistics we cannot say much about 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 2.23. *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 2.24. *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

Statistical results

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 2.25. *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 2.26. *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.

2.6 Discussion and conclusion

The results of the analysis of Own score are summarized in Table 2.27a and the results of the analysis of Percentile are summarized in Table 2.27b.

Table 2.27a. *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 2.27b. *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.
4. Absolute self-assessment is more responsive to information than relative self-assessment.
5. Although the simple model proposed in Chapter 1 of this dissertation 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 (see Chapter 1 of this dissertation) 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 Chapter 1 of this dissertation 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, however the difference was not significant. Based on this observation and from visual inspection of the results, the miscalibration in relative measure seems to be, based on 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[61] 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.

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

Table 2.28. *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 2.29. *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 2.30. *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 2.31. *Overestimation in Percentile in Experiments 2 and 3 by quartiles.*[62]

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 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)[63], or by the difference in the tasks alone. To

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

[63] 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

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.[64] Moreover, the simple model of ability perception (see Chapter 1 of this dissertation) might be extended.

happen in Experiment 1, where the incentives for as good performance as possible are much higher (admission to CERGE-EI).

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

References

Brueggen, A., Strobel, M., 2008. Real Effort versus Chosen Effort in Experiments. *Economics Letters, 96 (2),* 232-236.

Burson, A.K., Larrick, P.R., and 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., and Johannesson, M., 2006. Confidence Interval Estimation Tasks and the Economics of Overconfidence. *Journal of Economic Behavior and Organization, 61 (3)*, 453-470.

Dhami, K.M., Hertwig, R., and 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., Dunning, D., and Kruger, J., 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., and Rutstroem, E.E., 2005. Characterizing the Entrepreneur Using Field Experiments. *Working Paper 05-30, Department of Economics, College of Business Administration, University of Central Florida*.

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., and 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., and 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., and 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., and 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.

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.

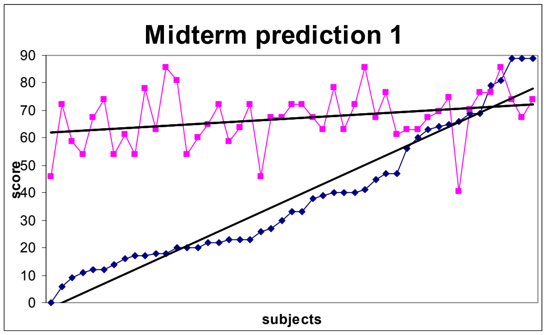
Rydval, O., Ortmann, A., 2004. How Financial Incentives and Cognitive Abilities Affect Task Performance in Laboratory Settings: An Illustration. *Economics Letters, 85 (3),* 315-320.

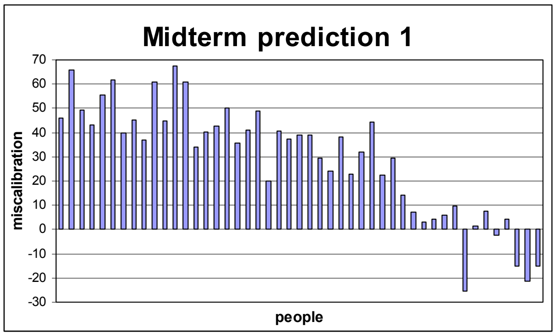
Smith, V.L., 2002. Method in Experiment: Rhetoric and Reality. *Experimental Economics, 5 (2),* 91-110.

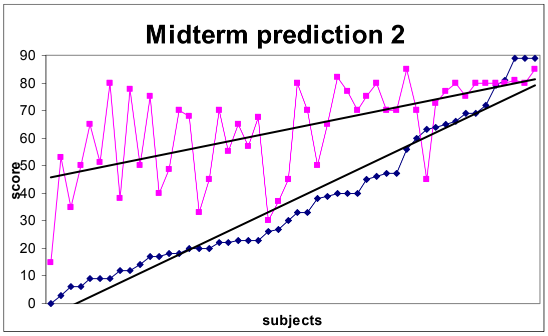
2.7 Appendix

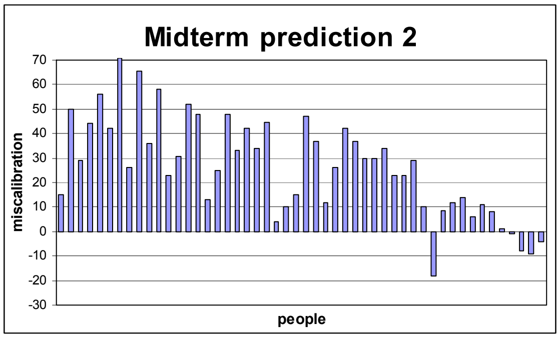
Appendix A

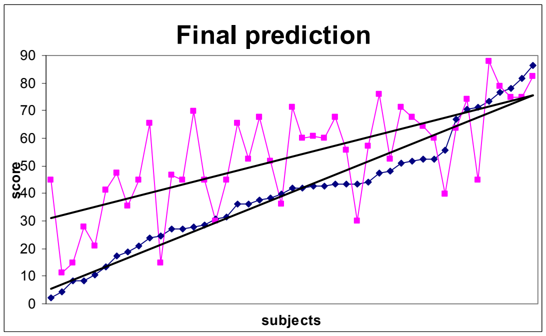
On the left, graphs of the real (blue line) and estimated (pink line) distribution of score (together with the corresponding trend lines). On the right, graphs of distribution of miscalibration of Own score (all data adjusted and ordered from the lowest to the highest real score).

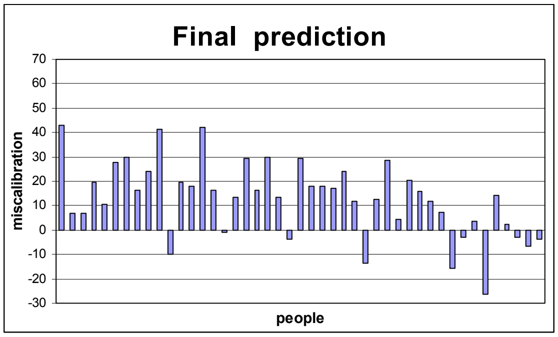




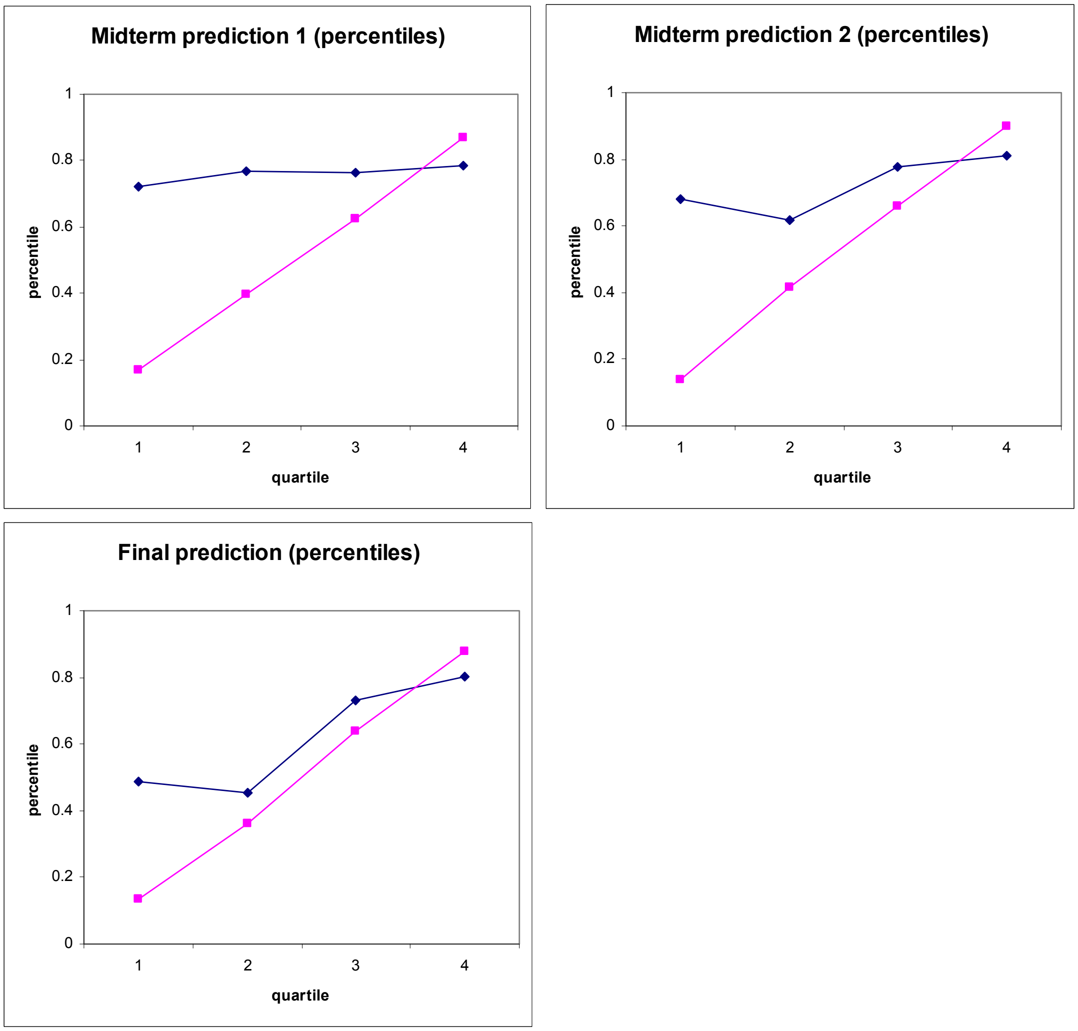








Graphs of actual percentiles (pink line) and estimated percentiles (blue line).



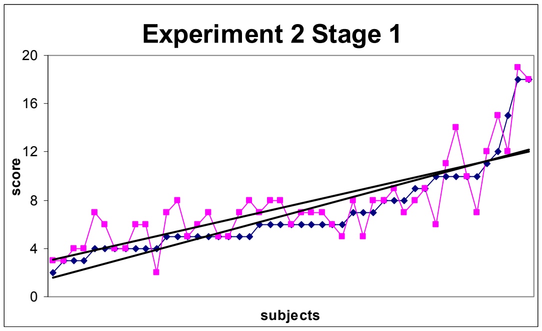
Midterm prediction 1 (percentiles)

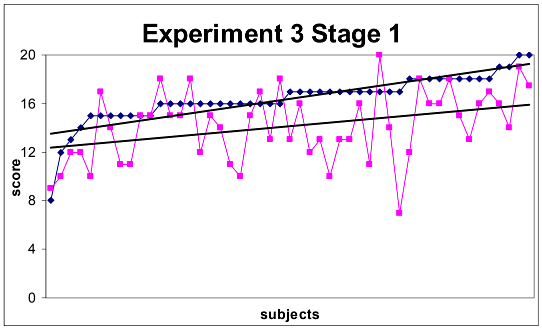
Midterm prediction 2 (percentiles)

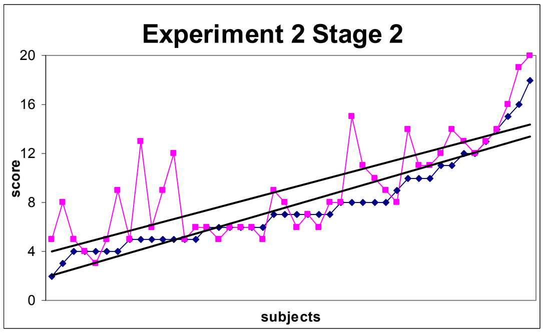
Final prediction (percentiles)

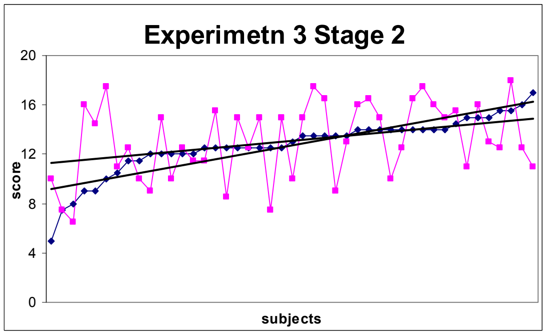
Appendix B

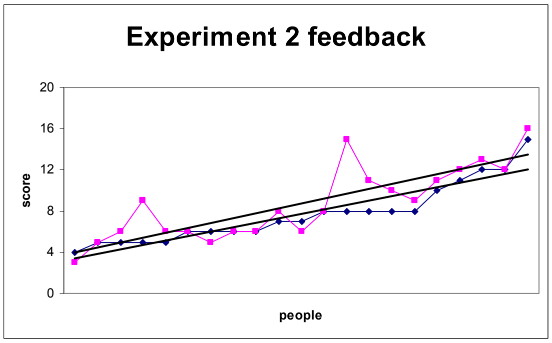
Graphs of the real (blue line) and estimated (pink line) distribution of score (all data adjusted and ordered from the lowest real score to the highest real score); together with trend lines for each series.

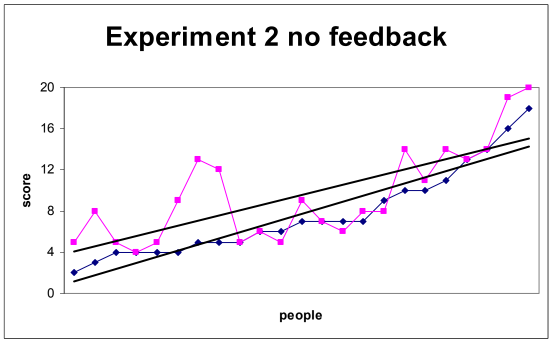

Experiment 2 Stage 1

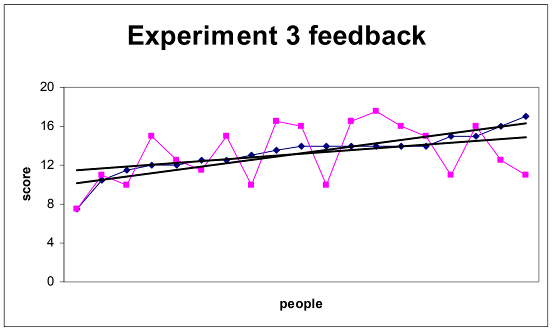

Experiment 3 Stage 1

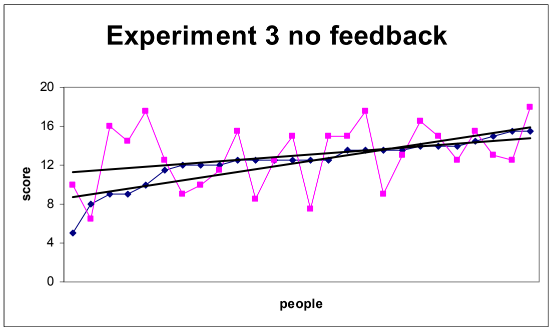

Experiment 2 Stage 2


Experimetn 3 Stage 2

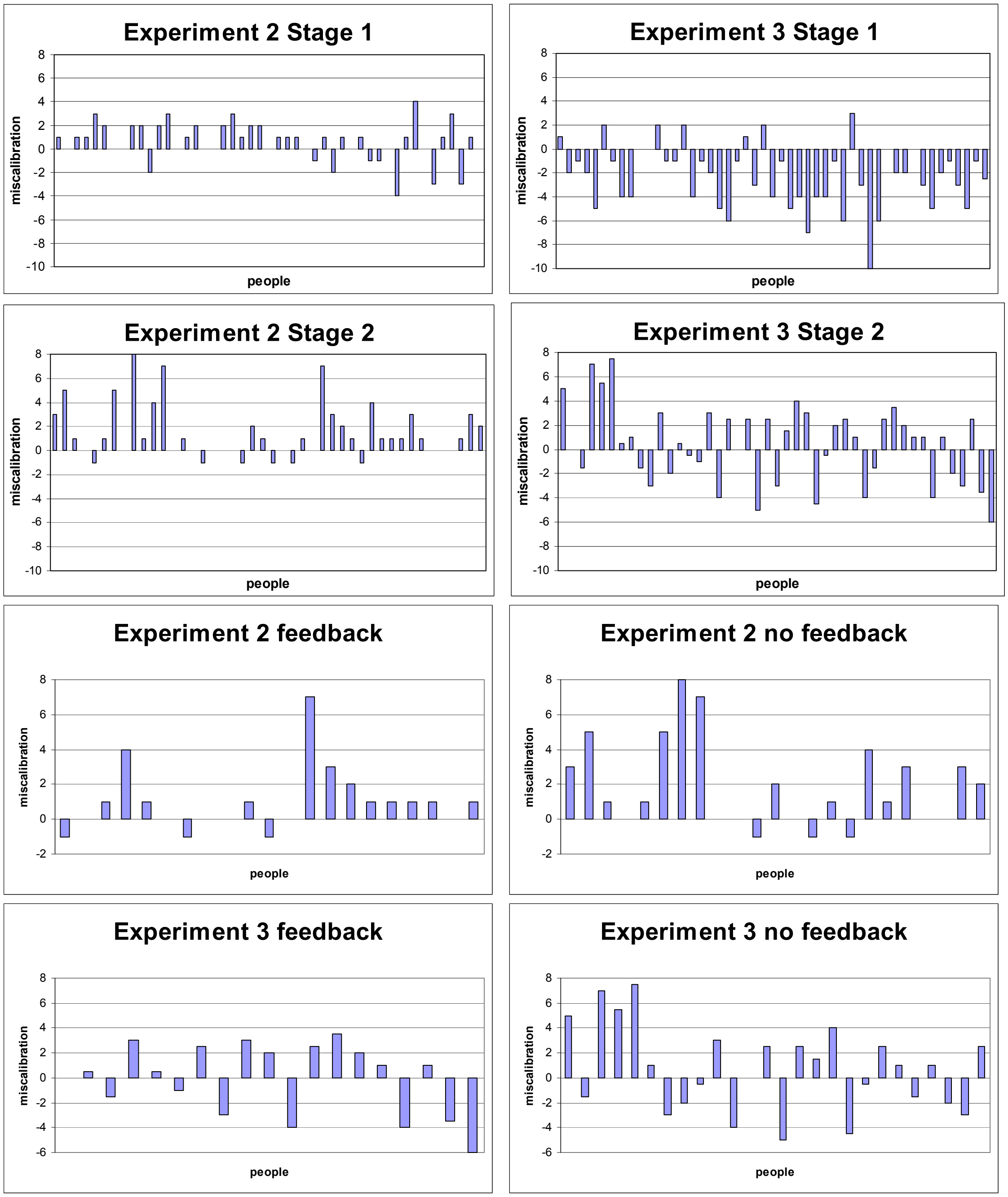

Experiment 2 feedback


Experiment 2 no feedback

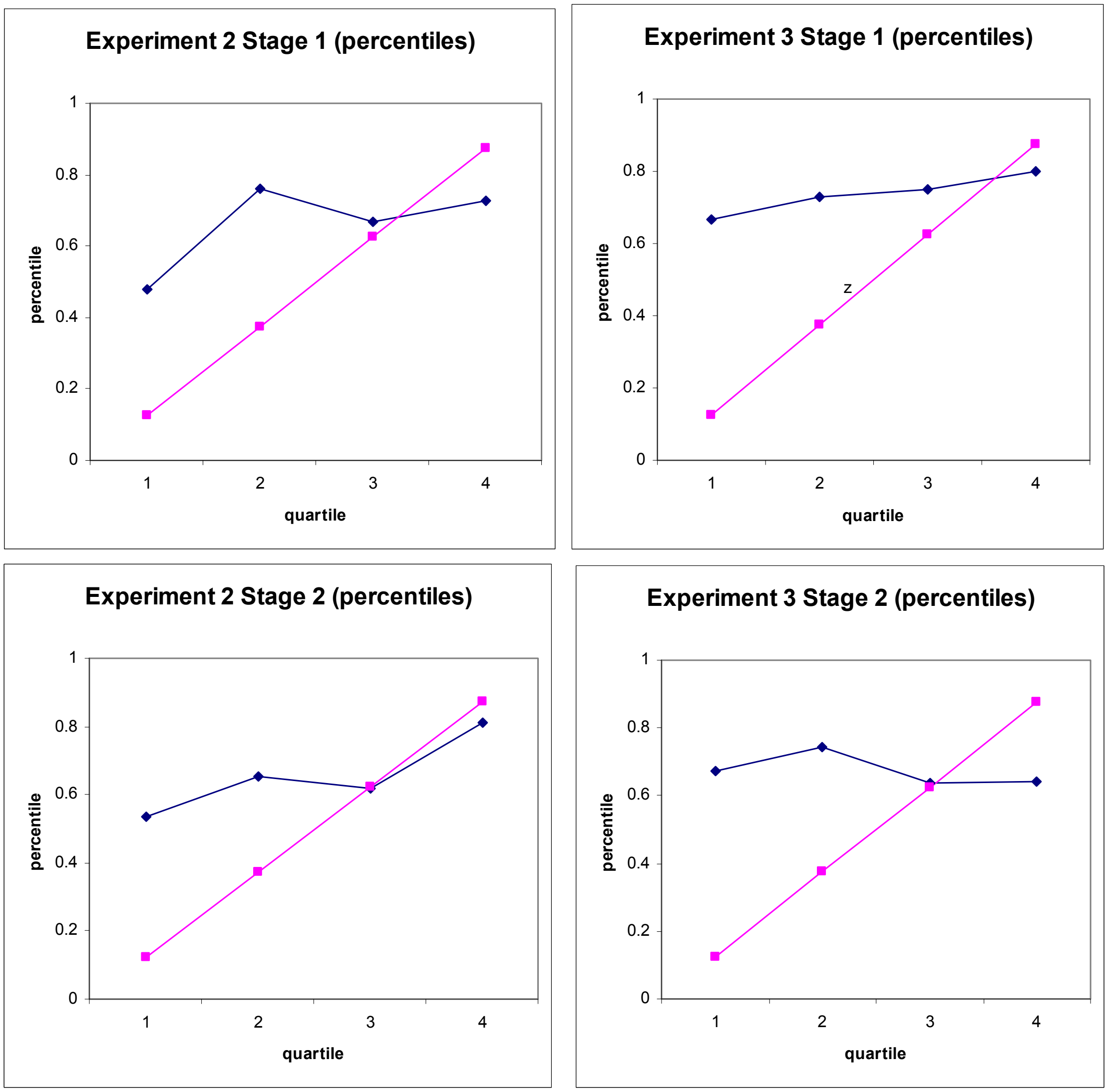

Experiment 3 feedback


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 C

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)[65].

Table 2.32. *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 D

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 2.32 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.

[65] 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.

CHAPTER 3

Overconfidence in Business, Economics, Finance, and Psychology

Abstract

Psychologists seem to have been the first to have claimed that overconfidence is a widespread phenomenon. For that reason, we first review the main research results from psychology and identify the main methodological issues there. We then briefly review the main issues identified so far in economics and find that overconfidence studies in business, economics, and finance are more diverse than those in psychology. We identify nine paradigms, categorize experimental studies from the business, economics, and finance literatures and review them for each paradigm separately. We pay attention to methodological issues identified in psychology as well as in economics and point out the main shortcomings of the reviewed studies. We end with suggestions for further research.

3.1 Introduction

Many mainstream economic theories explain the functioning of financial markets (e.g., Cochrane, 2001; Fama, 1998; Milne, 2003) and economic behavior of individuals (e.g., Eggertsson, 1990; Rosenthal and Rosnow, 2006). According to these theories, people are rational decision makers that induce markets to be reasonably efficient. However, the empirical reality inside and outside the laboratory does not always seem to support the efficient market hypothesis, or economists' assumptions about people's decision-making abilities. A barrage of choice "anomalies" has been identified over the past decade (e.g., Camerer, 1995, 2003; Camerer et al., 2003; Friedman, 1998; Hirshleifer, 2001; Shiller, 2003; Shleifer, 2000; Stracca, 2004) that seem to question the assumption of rationally acting agents in efficient markets. According to Stracca (2004), for example, anomalies frequently observed in financial markets include decision heuristics, ambiguity aversion, overconfidence, narrow framing, disposition effect, endowment effect, and preference reversal. Krueger and Funder (2004) list some additional errors of judgment identified and labeled by social psychologists (e.g., false consensus effect, confirmation bias, self-serving bias, hindsight bias, halo effect).

Overconfidence is one of the most frequently and widely investigated anomalies. It has been a hot topic in the business literature (e.g., Malmendier and Tate, 2005; Koellinger, Minniti, and Schade; 2007), economics (e.g., Camerer and Lovallo, 1999; Noeth and Weber, 2003; Van den Steen, 2004; Compte and Postlewaite, 2004; Vigna and Mamendier, 2006), finance (e.g., Barber and Odean, 2000, 2001; Daniel, Hirshleifer, and Subrahmanyam, 2001; Hirshleifer, 2001; Chuang and Lee, 2006; Menkhoff, Schmidt, and Brozynski, 2006), and other related literatures (e.g., marketing – Alba and Hutchinson, 2000; Meloy, 2000; Gershoff and Johar, 2006; health – Armantier, 2003). Even though overconfidence has recently been a hot topic in business, economics and finance, it has been a topic of interest to psychologists even longer.[1]

The findings on overconfidence, experimental as well as empirical, are ambiguous. In this review, we concentrate on the experimental literature. While the number of papers identifying the overconfidence (or underconfidence) bias has increased at a steady pace, studies questioning its existence have also emerged (e.g., Fama, 1998; Dhami, Hertwig, and Hoffrage, 2004; Gigerenzer, Hoffrage, 1995; Juslin, Winman, and Olsson, 2000; Krueger and Mueller, 2002; Krueger and Funder, 2004). Some of these studies point out methodological problems with the definitions of biases (e.g., Juslin, Winman, and Olsson, 2000; Dhami, Hertwig, and Hoffrage, 2004), others question the ecological validity of the experimental environments (e.g., Dhami, Hertwig, and Hoffrage, 2004; Gigerenzer, Hoffrage, Kleinboelting, 1991), and some suggest an alternative (non-bias) explanation of the seemingly well-established results (see e.g., Chapter 1 of this dissertation). In light of this emerging evidence it seems safe to say that the overconfidence bias is not universal for all domains and environment. Specifically, the question whether overconfidence

[1] For a review of an important subset of the experimental literature in psychology see e.g., Juslin, Winman, and Olsson (2000).

exists in business, economics, and/or finance environments (and if yes then under which circumstances and in what magnitude) has not been answered in a satisfying manner.

Better understanding of the problem of overconfidence would help researchers to better model people's behavior in various situations and domains. For example, investigating and clarifying the behavior of entrepreneurs would facilitate our understanding of market-entry decisions and, possibly, help entrepreneurs to avoid unfortunate choices. Understanding (the reality of) overconfidence might also, through the design of appropriate tutorials, help the general public to improve performance on activities such as driving. In fact, (the reality of) overconfidence is important in all situations where decisions (be they economic or other decisions) are made based on any subjective assessment (absolute or relative).

The goal of this literature review is to categorize and evaluate, based on a reasonable and replicable selection mechanism, experimental studies on overconfidence from economics, business, and finance. For each category, we identify the main design and implementation problems, other flaws, alternative explanations and suggestions for further research. By doing so, we utilize findings from other fields (especially psychology).

The remainder of this chapter is organized as follows: In the second section we review the basic findings on overconfidence from psychology and point out the most important contributions to overconfidence research. Section 3.3 briefly illustrates the diversity of the paradigms used in business, economics, and finance. In Section 3.4 we identify the most widely used paradigms used in business, economics, and finance (namely, general-knowledge questions, confidence intervals, forecasting, market-entry games, auctions, willingness to sell/buy, information, assessment of others, and self-awareness questions) and we then review the relevant studies for each paradigm separately, concentrating on the major methodological issues described in Sections 3.2 and 3.3. Section 3.5 concludes.

3.2 Psychology

Overconfidence (as well as its twin sibling – underconfidence) has been a perennial topic in psychology for decades. As the literature reveals, over- and underconfidence emerge primarily in calibration studies where they are a result of types of self-assessment in various situations. Two fundamental types of self-assessment can be usefully distinguished: absolute and relative. In absolute self-assessment, overconfidence (underconfidence) usually occurs in the case of a positive (negative) difference between estimated/predicted and observed probabilities or frequencies for some events or statements. In addition, formation of too narrow (wide) confidence intervals is also often interpreted as overconfident (underconfident) behavior. In relative self-assessment, overconfidence (underconfidence) bias is identified when one claims to be better (worse) ranked within a group than he/she really is.

The most widely used instrument of absolute self-assessment is general-knowledge questions. Therefore, in this section we will focus on utilizing such questions. We then review the research on calibration in sensory discrimination tasks. We also review personal assessment (self-awareness) questions that seem to be the most widely used instrument to determine people's relative self-assessment (note that general-knowledge questions as well as any other task can be used as the underlying task for personal assessment questions). Finally, we summarize the main issues identified so far in psychology that might (and actually do) play a role in business, economics, and finance.

3.2.1 General-knowledge questions

Arguably the most often used paradigm in psychology is general-knowledge questions. These are pairs of questions such as "Which country has the larger population: (a) Sweden or (b) Mali?" followed by "How confident are you about your answer? 50% 60% 70% 80% 90% 100%?"[2] Overconfidence (underconfidence) is defined as the mean subjective probability, or confidence, being greater (smaller) than the proportion of correct answers. In studies from psychology, overconfidence is a more common result than underconfidence. The following studies cover the basic conceptualizations and directions of research on general-knowledge questions in psychology.

Erev, Wallsten, and Budescu (1994), motivated by the contradictory over- and underconfidence experimental results of previous research, pointed out that two basic response modes were used in general-knowledge questions research. In the first response mode, used in the revision-of-opinion literature, "prior probabilities of hypotheses and conditional probabilities of data are well-defined stochastic properties of the environment" (p. 520). These kinds of studies on Bayesian updating were produced primarily in the 1960s and the results exhibit underconfidence. The second response mode, used in the calibration literature, operated with paradigms "in which uncertainty is due primarily to lack of knowledge rather than to well-defined environmental factors" (p. 520). This stream of literature continues to be published and typically exhibits overconfident behavior.

The importance of the regression-to-the-mean – a statistical artifact which occurs when one has two imperfectly correlated measures and causes the measured ability to regress toward the mean – in the research on overconfidence was introduced in Pfeifer (1994). The author created an idealized calibration study to show that under imperfect knowledge even well-calibrated individuals can generate a calibration curve[3] with slope less than unity (seemingly capturing miscalibration). The results, computed based on probability theory, showed convincingly that the calibration curve was flatter for non-experts (this corresponds to the previously observed better calibration of experts). The calibration curve was also flatter for difficult tasks (meaning means worse calibration for difficult

[2] This is called a half-range response mode. The second response mode is called a full-range response mode – to state confidence in the correctness of a statement, e.g. "How confident are you in the correctness of the following statement? Sweden has a larger population than Mali." 0% (false) - 100% (true).

[3] The calibration curve depicts the actual proportion of occurrences of the event under consideration (on the vertical axis) versus subjective probability (on the horizontal axis). The calibration curve of a well-calibrated individual is the straight unity line.

tasks). The results of the idealized experiment suggested that subjects' probability estimates appear miscalibrated because of the regression-to-the-mean phenomenon. The regression-to-the-mean argument was later used also in the case of calibration in relative self-assessment (e.g., Krueger and Mueller, 2002).

In addition, Erev et al. (1994) created, based on a similar concept as Pfeifer (1994), a general model that assumed that subjects' confidence was a function of true judgments and error. A monotonic response rule translated subjects' confidence into an overt response. This model generated (through simulations) patterns similar to those that researchers obtained from the experiments and thus pointed out the role of error in judgment processes. In follow-up work, Budescu, Erev, and Wallsten (1997) generalized this model and underlined that there was a need for further investigation of overconfidence to determine whether it is purely a statistical artifact or a real bias. Budescu, Wallsten, and Au (1997) provided a method to determine whether the observed overconfidence was caused artificially by the random error or by the real effect of overconfidence. Their experimental results indicated that actual observed overconfidence was higher than the model with random error (Budescu et al., 1997) would predict for an unbiased judge. Even though these authors showed that part of overconfidence remains present after controlling for error, they clearly indicated the importance of random error in the judgment process (because it explains part of alleged overconfidence). Note that error in the judgment process relates to the regression-to-the-mean in the sense that a variable measured with an error can be regarded as an imperfect measure of the real variable. And this is the case when regression-to-the-mean occurs.

As we mentioned above, studies using general-knowledge questions suggested mostly overconfident behavior. However, several studies in the calibration literature reported contradictory results (over- and underconfidence) under seemingly similar conditions. The finding that overconfidence was observed for more difficult tasks and underconfidence for easier tasks established a new phenomenon, the so-called hard-easy effect (dependence of the bias on task difficulty). For example, Lichtenstein and Fischhoff (1977) showed experimentally that people's probability judgments were, despite their moderately good calibration, prone to systematic biases (where the most common bias was overconfidence) and that calibration depended on task difficulty.

Gigerenzer, Hoffrage, and Kleinboelting (1991) proposed a different explanation of overconfidence. The authors created a theory of probabilistic mental models (PPM) which was able to explain the overconfidence effect as well as the hard-easy effect. Moreover, this theory stressed out the confidence-frequency effect – the systematic difference between a judgment of the frequency of correct answers in the long run and a confidence judgment in a single event. The authors claimed that, when people face two-alternative general-knowledge questions, they first use a local mental model (LMM) according to which people try to determine the correct answer either by knowing the answer to the question or by process of eliminating the alternatives. If use of the LLM does not determine an answer then people use the PPM, which is based on the connection of the task-specific structure and a probability structure of the corresponding natural environment. In this model, people use several cues (e.g., that a German city has a

Bundesliga team, if one is asked to compare the number of inhabitants of two German cities) with various cue validities (0.91 for Bundesliga team) in a given reference class (all German cities with a population greater than 100,000). The model assumes that by using a cue, people choose the more probable answer and use the cue validity as a proxy for their confidence. The authors claimed that overconfidence (or underconfidence) occurs if unrepresentative sets of questions (alternatives) are used (because the cue validities do not correspond to the ecological validities). Thus, using representative sampling (e.g., of cities) from the reference class (e.g., German cities with population greater than 100,000) the theory predicted well-calibrated individuals. Moreover, Gigerenzer et al. (1991) pointed out the difference between confidence in single events and confidence in the long run. The confidence-frequency effect is, according to these authors, based on the fact that people use different reference classes when evaluating the probability of correctness of a particular answer (in this case, the reference class was e.g. all German cities with more than 100,000 people) and when they make judgments of the frequency of correct answers in a set of questions (here the reference class was sets of general-knowledge questions experienced in the past). Since the two reference classes differed (as did the cue validities), the model predicted calibration if a representative set of questions from that particular reference class was chosen (class of German cities with more than 100,000 people or class of general-knowledge questions). To support this theory, Gigerenzer et al. (1991) conducted two experiments and showed that the results were in line with PPM theory: The results on over- and underconfidence depended on the type of set of questions (representative or selected) and on the response format (confidence in each question or frequency of correct answers in the whole set).

Subsequently, Soll (1996) extended the PPM model by incorporating a random error component. The author claimed that random error could influence judgment in two ways: "in forming subjective feeling of confidence and translating it into external report" (p. 120) and "because subject had experience only with a small subset of environment…the validity of a set of cues within a given subset of the environment will likely differ from ecological validity" (p. 120). He assumed the reported confidence to be a function of ecological validity, bias, and error. To show the interaction between random error in judgment and the environment, Soll (1996) constructed four different question sets: three representative and one unrepresentative set. The author showed that the cues were less valid in the unrepresentative set than in natural ecology, which led to overestimation because subjects counted with higher cue validity than it actually was. Overconfidence was more distinct in the representative hard set than in the easy set. In addition, adding trick questions (unrepresentative set) caused a reduction in accuracy and an increase in overconfidence. Furthermore, Soll (1996) showed that random error in judgment can produce the hard-easy effect even in representative sets. The results of an experiment, which used the 50 most populous US cities without the use of their real names and four available cues (major league baseball, large airport, tall building, and length of daily commute), supported the results of the model.

Even though the results from psychology on overconfidence in general-knowledge questions are ambiguous and the theories described above do not completely rule out the presence of over- and underconfidence, psychologists Juslin, Winman, and Olsson (2000) offered a very interesting insight. The authors examined studies involving the overconfidence phenomenon and the hard-easy effect in two alternative general-knowledge questions. In these studies, the second response mode as described in Erev et al. (1994) is used – subjects are asked to estimate the probability that their answers are correct. The notion of overconfidence in the analyzed studies represents the observation that "the mean subjective probability assigned to the correctness to general items tends to exceed the proportion of correct answers" (p. 384). The "hard-easy effect" indicates "the covariation between over/underconfidence and task difficulty" (p. 384) (where overconfidence appears in hard item samples and underconfidence in easy item samples).

First, Juslin et al. (2000) discussed three methodological problems connected with the hard-easy effect (and ignored in most of the experimental studies): scale-end effect, linear dependency, and regression effect. All these three effects stem from the definition of overconfidence. Because of the restraints on the boundaries (subjective probability is defined only between 0.5 and 1.0, i.e., random choice and certainty, respectively) any fitted function will have a zero or negative slope. Similarly, a linear dependency between the proportion of correct answers and over/underconfidence can, according to Juslin et al. (2000), contribute to the hard-easy effect. Finally, because of a purely correlative relationship between objective and subjective probabilities, the authors argued that the regression effect contributed to overconfidence (especially when the proportion of correct answers was low).

Second, and most important, in their meta-analysis Juslin et al. (2000) inspected 130 experimental studies involving general-knowledge questions that satisfied the following conditions: two-alternative forced choice item, tests of general knowledge, no statistically significant effect of any independent variable on data, and over/underconfidence scores and proportion correct available. The authors identified 95 studies with selected samples and 35 studies with representative samples. Inspecting these data and controlling for the methodological problems, Juslin et al. (2000) identified almost zero overconfidence (and independent of task difficulty) in samples with representative items, whereas they found substantial overconfidence in samples with selected items. These results suggested that overconfidence as well as the hard-easy effect in two-alternative general-knowledge questions is an experimental artifact where the representativeness of stimuli (primarily sample selection in this case) plays the key role.

Apart from identifying the methodological deficiencies of current approaches, some researchers have tried to find ways to reduce overconfidence. For example, Arkes, Christensen, Lai, and Blumer (1987) attempted to reduce subjects' overconfident behavior with feedback on performance in the first five questions. Feedback caused subjects to be indeed less overconfident (even slightly underconfident) in the subsequent 30 questions. Similarly, the set of people who discussed their answers in a group later expressed less overconfidence. In addition, Pulford and Colman (1997) combined feedback and item difficulty in their experiments in order to de-bias their subjects. Overconfidence decreased with decreasing task difficulty resulting in underconfidence for easy questions and women were much less overconfident than men. Feedback had a significant effect on reducing overconfidence only for hard questions. Moreover, Koriat et al. (1980) showed that overconfidence bias can also be reduced by requiring subjects to consider reasons why they may be wrong.

To sum up, the more frequent finding in general-knowledge questions in psychology is overconfidence; however, the direction of overconfidence/underconfidence seems to depend on the response mode (Erev et al., 1994). Dependence of overconfidence on other factors was later extended by results that indicated the role of task difficulty – these results became known as the hard-easy effect (e.g., Griffin and Tversky, 1992). Some researchers pointed out methodological problems (e.g., regression-to-the-mean, error in the judgmental process) that might artificially contribute to overconfidence bias (e.g., Pfeifer, 1994; Erev et al., 1994). Most prominently, Juslin et al. (2000) showed by way of a meta-analysis that the hard-easy effect (and thus over- and underconfidence) in general-knowledge questions is likely to be an experimental artifact – it does not appear in studies where representative stimuli (random sample selection) are used. Finally, several ways to increase calibration were identified – feedback (Arkes et al., 1987), discussion of choices/results (Pulford and Colman, 1997), and counter-reasoning (Koriat et al., 1980).

3.2.2 Sensory discrimination

Sensory discrimination tasks are, in contrast to general-knowledge questions, non-cognitive tasks. Generally, in sensory discrimination tasks people are required to use some of their senses to judge what object better satisfies the given criterion. These tasks most frequently include visual comparisons of line-length of pairs of presented objects; but other types of measures have also been used (e.g., haptic[4] comparisons of weight). In contrast to general-knowledge questions, in experimental studies involving sensory discrimination tasks underconfident behavior prevails. In the subsequent paragraphs, we review the basic literature on sensory discrimination.

Bjoerkman, Juslin, and Winman (1993) experimentally identified underconfidence in sensory discrimination tasks. The authors did not find, even after 160 trials, any effect of feedback on calibration. They also developed subjective distance theory and showed that both their experimental results fit this theory. Later, Olsson and Winman (1996) argued that overconfidence in Baranski and Petrusic (1994) was achieved only by misleading subjects, who were using an unwarranted symmetry assumption. Olsson and Winman (1996) experimentally showed that it is very difficult to achieve overconfidence in sensory discrimination (even if the proportion of correct answers is low). The authors pointed out the difference in confidence (especially in its accuracy) in sensory tasks and cognitive tasks. In contrast, comparison of calibration in the reanalysis of data from Baranski and Petrusic (1994) with calibration in a new experiment in Baranski and Petrusic (1999) suggested that there might be cross-national differences in calibration in sensory discrimination tasks – underconfidence (overconfidence) for easy (hard) sensory judgements in Canada and a unique bias – underconfidence in Sweden.

[4] People are presented with two objects which they can weigh in their hands and have to decide which object is heavier.

As an explanation for the difference in calibration in sensory discrimination tasks and cognitive tasks, Juslin and Olsson (1997) proposed that there are two different modes of uncertainty in confidence judgments: Thurstonian (in sensory discrimination tasks with pair comparison) and Brunswikian (in cognitive tasks). The first mode of uncertainty involves an error where “information provided to the sensory transducers allows error-free performance, and erroneous decisions arise because of the limited reliability of the sensory transducers” (p. 345).[5] In the second uncertainty mode, “erroneous decisions arise due to the less-than-perfect correlations between known aspects (cues) and unknown current or future aspects or states of the world” (p. 345). The authors also presented a model that fits the patterns of the experimental results. Olsson and Juslin (2000) further explained the difference between Brunswikian and Thurstonian error and presented a sensory sampling model. The authors showed that their model captures the experimental findings in sensory discrimination tasks (underconfidence that is unaffected by feedback or perceptual illusions).

Later, Juslin, Winman, and Olsson (2003) directly contrasted (theoretically and experimentally) both task types – general knowledge and sensory discrimination. The experimental data fit the theoretical predictions; particularly, for full-range probability assessments in general-knowledge tasks the authors reported good calibration, and for full-range probability assessments in sensory discrimination tasks they found underconfidence; however, Juslin et al. (2003) identified much higher overconfidence for confidence intervals than the model predicted.

The literature suggests that, unlike general-knowledge tasks, sensory-discrimination tasks tend to produce underconfident behavior that seems not to improve with feedback (Bjoerkman et al., 1993). The importance of representative stimuli also turned out to be important in sensory discrimination tasks (Olsson and Winman, 1996). The substantial difference in calibration in general-knowledge questions and in sensory-discrimination tasks seems to be caused by the different types of error involved in the confidence judgements of these two paradigms (Juslin and Olsson, 1997).

3.2.3 Personal assessment (self-awareness) questions

In Sections 2.1 and 2.2, we reviewed literature which deals exclusively with absolute self-assessment. In psychology, there are many studies that also deal with relative self-assessment in various domains. The paradigm used in these studies is personal assessment questions and it compares people’s subjective evaluation relative to others. In many cases, self-assessments in absolute and relative measures are closely connected.

Well documented is the better-than-average (BTA) or above-average effect – the human tendency to overestimate one's achievements and capabilities in relation to others (e.g., driving abilities – Svenson, 1981). However, later it was shown that people may also exhibit a worse-than-average (WTA) or below-average effect in tasks that are difficult or

[5] This means that unmodified information about the object under investigation (e.g., the length of a line) gets to the sensory transducer (the eye). The error then emerges on the way from the eye to the brain.

where success is rare (e.g., Kruger, 1999). Moore (2007) reviewed the literature on the BTA and WTA effect and concluded that BTA is usually found in studies involving easy (common) tasks while WTA in studies with difficult (rare) tasks.

Kruger and Dunning (1999) and Dunning et al. (2003) extended the elementary observation of the BTA or WTA effect and introduced the unskilled-and-unaware problem, which combines both over- and underconfidence. The authors suggested that, across many intellectual and social domains, the subjects that perform worse (the "unskilled") also lack the meta-cognition that allows them to assess their deficiencies. The experimental results also suggested that the very skilled subjects are, but less so, unaware of their skills. The authors explain this behavior by overconfidence of the unskilled and underconfidence of the very skilled.

The original findings (identified through test of students on grammar, logical reasoning, and humor) have since been replicated with different tasks (e.g., Dunning et al., 2003: classroom exams) and also different subject pools (Parikh et al., 2001: medical students assessing their interview skills; Edwards et al., 2003: clerks evaluating their performance; Haun et al., 2000: medical lab technicians evaluating their on-the-job expertise).

These results have attracted critical attention. For example, Krueger and Mueller (2002) argued that the results were caused by the so-called regression-to-the-mean – a statistical artifact which occurs when one has two imperfectly correlated measures (such as abilities and the perception of abilities) and causes the measured ability to regress toward the mean, which induces overestimation (underestimation) of the bottom (top) performers.[6] Krueger and Mueller (2002) pointed out that in the presence of the better-than-average effect the bottom performers overestimate their performance even more, because then the real mean is higher and there is more space for the regression-to-the-mean among the unskilled. The authors also showed that the over- and underestimation disappears after controlling for the unreliability of measures and measurement errors. In response, Kruger and Dunning (2002) cast doubt on the results by Krueger and Mueller (2002), arguing that Krueger and Mueller (2002) used unreliable tests and inappropriate measures of the relevant mediating variables; Kruger and Dunning (2002) also pointed out that the results by Krueger and Mueller (2002) are valid only if low or moderate levels of reliability are used and not in samples with highly reliable measures.

The results in Burson, Larrick, and Klayman (2006) suggested that the unskilled-and-unaware problem depends on task difficulty; specifically, bottom performers are equally (or even better) calibrated than top performers for harder questions while they are calibrated worse for easier questions. Thus, based on these results, task difficulty seems to play an important role in relative self-assessment as in the case of general-knowledge questions in absolute self-assessment.

Moore and Healy (2008) employed the assumption that people have imperfect knowledge of their own performance and even more imperfect knowledge of others' performance.

[6] Therefore, because of the regression-to-the-mean, the expected perceived performance of bottom (top) performers shifts toward an average self-assessment.

Based on this assumption, the authors showed that simple Bayesian belief updating explains the negative relationship between overestimation (absolute self-assessment) and overplacement (relative self-assessment). The unskilled-and-unaware problem was further investigated in economics (see Section 3).

This paradigm – personal assessment questions – is more complicated than general-knowledge questions because it involves absolute and relative self-assessments and therefore it is more sensitive to various external factors. As in the case of general-knowledge questions, overconfident behavior is more frequent than underconfidence. In addition, the studies reviewed above highlighted the difference in calibration of the skilled and the unskilled – the unskilled-and-unaware problem (e.g., Kruger and Dunning, 1999). Personal assessment questions tie the overconfidence question to the question of calibration, Bayesian updating, information available (e.g., Moore and Healy, 2008), feedback, incentives, etc. All of these issues are known to be of importance to people's (not only) economic behavior.

In addition to the three paradigms that we reviewed above, Gigerenzer and Hoffrage (1995) described the importance of the information format. The authors showed that Bayesian algorithms were computationally simpler in frequency formats (information is acquired in natural sampling) than in probability formats. They supported this finding through an analysis of several thousand solutions to Bayesian problems where the subjects derived 50 % of all inferences by Bayesian algorithms and demonstrated that usage of natural frequencies helped to increase the number of correct answers. Attentional demand, they argue, is an important aspect of natural frequencies (because only two pieces of information are needed and rate base need not be attended to). Gigerenzer and Hoffrage (1999) experimentally showed that teaching frequency representations fosters insight into Bayesian reasoning. Because of some confusion among researchers, Gigerenzer and Hoffrage (1999) and Hoffrage, Gigerenzer, Krauss, and Martignon (2002) emphasized what natural frequencies are and what they are not (e.g., they are normalized frequencies). Therefore, researchers should pay attention not only to the quantity of information they provide but also to the quality and its format.

Probably the main insight from overconfidence studies in psychology is the representative design of the experiments. A case in point is Dhami, Hertwig, and Hoffrage (2004) who were concerned with representative design in experimental environments. The authors claimed that overconfidence in general-knowledge questions (Juslin, Winman, and Olsson, 2000) disappeared, and hindsight bias (Winman, 1997) was significantly reduced, when representative designs were used. Dhami et al. (2004) went on to claim that representative designs relate not only to representative sampling of participants and experimental stimuli but also to the response formats (e.g., natural frequencies seem to be more representative to subjects than probabilities; Gigerenzer and Hoffrage, 1995). In addition, feedback (e.g., Arkes et al., 1987) and counter-reasoning (e.g., Koriat et al., 1980) seem to play an important role in calibration studies.

3.3 Business, economics, and finance

There is a growing body of literature dealing with overconfidence bias in business, economics, and finance. In the introduction we cited some of the most prominent studies in these fields. To illustrate the heterogeneity of contexts and results, we first briefly describe these studies followed by a review of the main issues influencing the results in business, economics, and finance in order to sketch out the problems when reviewing the experimental studies in Section 4.

Managers allegedly overestimate the returns to their investment projects and find external funds too costly (Malmendier and Tate, 2005), individuals starting a business relied much more on their own perception than on objective evaluations (Koellinger, Minniti, and Schade, 2007), and overconfidence causes excessive business entry (Camerer and Lovallo, 1999). Noeth and Weber (2003) experimentally demonstrated that overconfidence has a negative effect on welfare, yet Compte and Postlewaite (2004) modeled situations depending on emotions and showed that in such a setup biases increased welfare. Cesarini, Sandewall, and Johannesson (2006) showed that miscalibration in interval estimation tasks is heavily dependent on the response format and that overconfidence can be reduced by pure repetition. Della Vigna and Mamendier (2006) illustrated that people going to a gym were in general overconfident about their future attendance and subscription cancellation probability. Barber and Odean (2000, 2001) demonstrated that excessively trading individuals earned less on average, that men traded more than women and that this excessive trading decreased men's earnings more. Chuang and Lee (2006) empirically found evidence that investors were overconfident about private information and traded more in risky securities; gains caused them to trade more aggressively which contributes to excessive trading volatility. Menkhoff, Schmidt, and Brozynski (2006) empirically showed that herding behavior of fund managers decreased with experience but their results did not exhibit a clear effect of experience on overconfidence and risk taking.

Alba and Hutchinson (2000) described several studies dealing with calibration of consumers. They concluded that high levels of calibration are achieved rarely, while moderate levels are more common. Interestingly, accuracy and confidence were in some cases absolutely uncorrelated, suggesting that either people have no idea about their performance or they do not base their confidence estimates on accuracy. Meloy (2000) experimentally showed that creating a good mood doubled the biased evaluation of new product information. In addition, information disconfirming tentative preference for a brand (favoring the other brand) had a much greater impact than negative information about the preferred brand. Finally, Armantier (2003) demonstrated that subjects making lethal risks estimates exhibited similar biases when estimating their own-age-cohort and entire population risk; concretely, overestimating rare risks and underestimating common risks. These results suggest that, compared to studies in psychology, studies on overconfidence in business, economics, and finance display a much higher variability in contexts and results.

In addition to the methodological issues discussed in psychology and reviewed in the previous section, economics research has addressed other important methodological issues. Perhaps the most interesting one for research on overconfidence in economics is that of incentives. Camerer and Hogarth (1999) investigated whether financial incentives in experiments matter or not. The authors reviewed 74 experiments with three different levels of performance-based incentives (no, low, and high). The results suggested that even though financial incentives lowered variance, the modal effect on performance was neutral. However, the authors found a significant improvement in performance in studies where judgment tasks and decision tasks were responsive to better effort. Incentives usually made no difference in tasks involving market trading, bargaining, or some risky choices. All studies where incentives hurt performance were from judgment and decision tasks. Nonetheless, as Hertwig and Ortmann (2001) noted, there are at least four reasons why economists do use financial incentives: they reduce performance variability, they are easier to implement than other incentives, non-satiation over the course of an experiment, and testing of economic theories (the majority of which are based on utility maximization). The authors concluded that financial incentives matter more in areas involving judgment and decision making and that financial incentives contribute to the data's convergence toward performance criterion and reduction of variance. Financial incentives thus are an important part of experimental design and researchers should carefully consider the effect of financial incentives on performance and use the appropriate incentives scheme in every experiment.

Harrison and List (2004) stressed the importance of the external validity of experiments. The authors proposed that the following six factors are of importance in this context: the nature of the subject pool, the nature of the information that the subjects bring to the task, the nature of the commodity, the nature of the task or trading rules applied, the nature of the stakes, and the environment that subjects operate in. When designing experiments, all these factors should be taken into account in order to get "useful" (externally valid) results.

Angner (2006) pointed out the importance of identifying and solving the overconfidence issue when he argued that in real life economists are, just as other subjects, victims of overconfidence. Yet economists' overconfidence could have dramatic consequences, as they often make high-confidence judgments and the tasks are frequently challenging – this is, according to previous research, associated with higher overconfidence. Overconfidence among economists could lead to misguided policy decisions and undermine trust in economics. Angner (2006) further questioned whether economists acting as experts realize they might be overconfident. One of the reasons to answer *no* is insufficient feedback. Finally, the author outlined solutions and remedies. Since one can assume that economists are doing their best to achieve highest possible performance (they have sufficient incentives), Angner (2006) suggested decreasing experts' confidence ratings to improve their calibration. Two methods might achieve this goal: requiring experts to provide arguments against their own view (to reason why they might be wrong) and providing frequent, prompt and unambiguous feedback. He also pointed out that some specialists are indeed extremely well calibrated (e.g., meteorologists).

In Chapter 1 of this dissertation we offered an alternative explanation of the above mentioned unskilled-and-unaware problem. We argued that the subject pools used by Dunning, Kruger, and their collaborators were not distributed uniformly, or at least symmetrically, but rather skewed toward the bottom. We constructed a simple model based on a J-distribution of abilities and the presence of unsystematic noise in the self-assessment process. The results of model simulations replicated the patterns identified in the experiments. We showed that the unskilled, rather than being more unaware than the skilled, face a tougher inference problem which, at least partially, explains their alleged lack of metacognitive ability. Thus, in Chapter 1 we showed that even well-calibrated, error-making agents can exhibit the observed pattern. We also demonstrated that the first assumption (J-distribution) can to some extent be weakened while the qualitative results remain the same. We also discussed the conditions under which they expect the unskilled-and-unaware problem to disappear. In follow-up work (see Chapter 2 of this dissertation), we conducted three experiments (one field and two laboratory) through which we identified a strong positive effect of general as well as specific information on calibration in relative self-assessment. In addition, the results suggested that it is the unskilled who improve their calibration most. We concluded that the unskilled-and-unaware problem arises mainly because of the lack of information and that feedback improves calibration in absolute self-assessment more than in relative self-assessment.

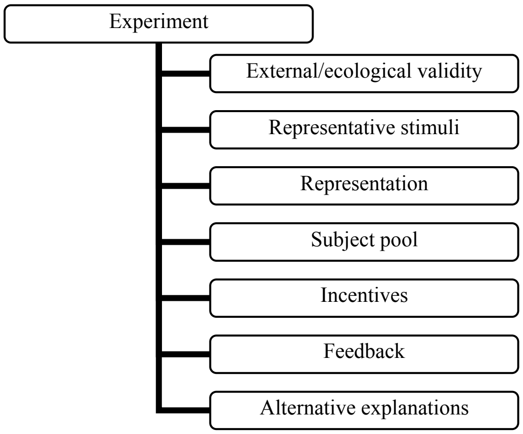
The most important issues identified so far in experimental studies in business, economics, and finance seem to be use of financial (or other) incentives (e.g., Camerer and Hogarth, 1999), external/ecological validity (e.g., Harrison and List, 2004), and use of appropriate subject pool (see Chapter 1 of this dissertation).

3.4 Paradigms

In this section we first summarize the most important issues identified in psychology as well as in business, economics, and finance. We then identify paradigms used in experimental studies on overconfidence in business, economics, and finance. Finally, we review and discuss relevant experimental studies according to various paradigms used in these studies while paying special attention to the issues identified in psychology as well as in business, economics, and finance.

The most important issues that we identified in sections 3.2 and 3.3 were the question of representativeness of stimuli (e.g., Juslin et al., 2000; Budescu et al., 1997; Gigerenzer et al., 1991; Soll, 1996) and external/ecological validity (e.g., Harrison and List, 2004), which are mostly an experimental design issue; the issue of experimental implementation: financial or social incentives (e.g., Camerer and Hogarth, 1999), representation issues (e.g., the usage of frequencies rather than probabilities – Brenner, Koehler, Liberman, and Tversky, 1996; response format dependency – Juslin, Wennerholm, and Olsson, 1999), and subject pools (see Chapter 1 of this dissertation); the issue of feedback – (e.g., Petrusic and Baranski, 1997). In addition, we also include possible alternative explanations of the results (e.g., risk aversion). Figure 3.1 lists these issues.

Figure 3.1. *Diagram of important issues in the reviews of experimental studies.*



As we already sketched above, there are many different paradigms used in studies on overconfidence in business, economics, and finance. We strove to create a non-opportunistic set of studies, i.e. a set selected from relevant sources using clear selection rules. The studies reviewed in this chapter were selected from the *Econlit* and *Web of Science* databases, as follows: First, we searched the databases for combinations of words {overconfiden*, underconfiden*, self-assessment} and {experiment*} "anywhere"[7] and created a corpus of texts. From this corpus we selected papers that were published in business, economic, or finance journals.[8] From the set we culled the experimental studies (excluding questionnaires, empirical studies, and theoretical models) dealing with the over-/underconfidence issue.[9] Altogether 45 studies satisfied our selection.

We identified the following nine different paradigms used in these experimental studies from business, economics, and finance:

- General-knowledge questions
- Confidence intervals
- Forecasting
- Market-entry games
- Auctions
- Willingness to sell/buy
- Information
- Assessment of others
- Self-awareness questions

[7] The search was performed on April 28, 2008. We also searched for "calibration", but this keyword returned too large a number of hits (mostly from other fields). Therefore we excluded it from the search.

[8] In the *Web of Science,* we filtered the studies according to journal category – including articles from journals categorized as "economics, business, finance, business finance, management". We also included working papers.

[9] We also excluded studies that did not investigate over-/underconfidence but only mentioned the word overconfidence (underconfidence) in the literature review, references, or in body of paper.

The process of defining the paradigms and assigning the studies according to the relevant paradigm was quite complicated. The paradigms vary substantially. Actually, we can lump the paradigms into two categories: elicitation method paradigms (Confidence intervals, Auctions, Willingness to sell/buy) and task domain paradigms (the remaining paradigms). Clearly, the paradigms are not mutually exclusive and some studies therefore fall under several.[10] We will review studies that can be classified under several paradigms in each relevant paradigm subsection.

Each study will be reviewed separately. We end (in italics) each review by concentrating on the important issues summarized in Figure 3.1 (except for studies, mostly from the 1990s, on the border between economics and psychology – from *Organizational Behavior and Human Decision Processes*, where we only briefly report the basic results). Whether and how the authors deal with these issues is summarized in tables in Appendix. We provide a more detailed discussion of the relevant issues (and other related literature) at the end of each paradigm subsection.

3.4.1 General-knowledge questions

General-knowledge questions are the most frequently used paradigm associated with over- and underconfidence bias in psychology. We reviewed the basic findings in Chapter 2 (Section 2.1). A series of papers published in *Organizational Behavior and Human Decision Processes* (a journal on the border between economics and psychology) concentrates on various aspects of overconfidence in general-knowledge questions. In the next few paragraphs we will briefly summarize the basic findings of these papers. Then we will review in more detail studies from business, economics, and finance that use general-knowledge questions or similar paradigms to investigate overconfidence. However, in this section we concentrate only on studies that use direct measures of confidence (half- and full-range confidence) and indirect measures (e.g., voting, lottery choices). Studies using confidence intervals to measure confidence will be reviewed separately because they require a different response format (range) and the results suggest significantly worse calibration.

Arkes, Christensen, Lai, and Blumer (1987) experimentally showed that feedback (on accuracy and discussion of the answers) significantly decreased overconfidence. The authors used in their experiments two alternative general-knowledge questions and the student subjects were asked to estimate their confidence in the correctness of each question using half-range confidence.

Subbotin (1996) studied the effect of feedback on confidence in related/unrelated and dependent/independent general-knowledge questions. In his experiments, students had to state half-range confidence in the correctness of their answers to two-alternative general-knowledge questions. Participants received feedback which consisted of the correct answer and of a reminder of the answer provided by the subject. The results suggested

[10] E.g., if confidence intervals are used to measure confidence in a forecasting study, the study will be classified under Confidence intervals and also under Forecasting.

that feedback, in the case of non-independent and related items, reduced underconfidence and improved calibration of underconfident judgments, yet feedback neither reduced overconfidence nor improved calibration of overconfident judgments. Furthermore, feedback significantly improved performance.

Schneider (1995) investigated the role of task difficulty and confidence-frequency effect (Gigerenzer, Hoffrage, and Kleinboelting, 1991) in a handwriting recognition task. Student subjects had to decide whether the handwriting sample had been written by a male or female. The results suggested that task discrimination decreased with the task difficulty. Furthermore, mean confidence judgments (half and full-range) exhibited slight overconfidence, while frequency estimates exhibited slight underconfidence.

Suantak, Bolger, and Ferrell (1996) examined the reasons for the presence of the hard-easy effect in general-knowledge questions. In their experiments, students had to express confidence (using half-range estimates) in their answers to general-knowledge questions. The results suggested that the ecological and ecological/random models (involving biased choice of stimulus material and random error in judgment) were not sufficient to explain the hard-easy effect.

Klayman, Soll, and Gonzáles-Vallejo (1999) concentrated on the relationship of overconfidence and task difficulty. As in the majority of studies, student participants estimated the correctness (half-range) of their answers on general-knowledge questions in various domains. The authors found that the degree of overconfidence (which was overall modest) varied over domains, yet was not a function of domain difficulty.

Juslin (1994) tested the impact of sample selection (random vs. selected) on calibration in general-knowledge questions. Student participants in the experiment were paid according to the time spent in the experiment. For each question, they had to choose the correct answer and state their confidence in that answer. The questions referred to comparisons of 164 countries. The results showed overconfidence in the selected questions treatment. However, overconfidence disappeared when a random selection of questions was used.

Soll (1996) developed a model and tested it experimentally. Students had to state which of two anonymous cities is more populous based only on the given information about these cities. The author showed that sampling error had a big effect on both calibration and overconfidence. Moreover, subjects appeared to be overconfident in the case of traditional measuring, yet well calibrated or slightly underconfident if reported confidence was compared to the objective probabilities – most subjects seemed to report validity of the information that they used accurately, but with substantial random error.

Stone (1994) experimentally demonstrated that positive expectations encourage overconfidence in choice accuracy. However, mildly negative expectations increased effort (compared to the strongly negative expectations), attention to strategy, and performance (number of correct answers). The authors concluded that in some tasks mildly negative expectations might have a better effect on performance than positive expectations. On the other hand, positive expectations increased overconfidence in choice accuracy.

Yates, Lee, and Bush (1997) inspected the dependence of overconfidence on the response styles of American and Chinese students. As in other studies, participants had to answer two alternative general-knowledge questions and state half-range confidence in the correctness of each answer. In addition, they could avoid participating in a wager (which would choose randomly one answer and pay money only if the corresponding answer was correct) or sell this possibility. This allowed the authors to compute inferred confidence. The results implied that Chinese students are more overconfident than American students when overconfidence is measured directly and almost equally overconfident when the inferred confidence is analyzed.

Price (1998) concentrated on response modes. Students in his experiments were asked to report their confidence (in the form of relative frequencies and probabilities) in the correctness of each answer to the general-knowledge questions. The results indicated that the responses in the relative-frequency mode were less dispersed and expressed certainty less than responses in the probability mode.

Juslin, Winman, and Olsson (2003) experimentally showed that people are well calibrated for general-knowledge questions if full-range confidence estimates are used, yet underconfident for sensory discrimination (they were asked to express their belief that the statement – general knowledge or comparison of the length of two lines – was true). The authors identified more deviations from additivity (sum of estimated probabilities for a statement and its negation) in sensory discrimination task than in general knowledge.

Recently, Sieck, Merkle, and Zandt (2007) tested the conjecture that overconfidence in two-alternative general-knowledge questions is partly caused by option fixation – "a tendency to evaluate only the high-familiar option." The authors conducted two experiments with students where they used various ways of stating confidence – a) confidence stated separately for each option independently and b) confidence stated for the selected option from two options (standard procedure). The results revealed that overconfidence decreased if the confidence statements for both options were assessed separately and if an explicit choice was prompted to be made. Sieck et al. (2007) also showed that people tend to fixate on one option. In the second experiment the authors showed that overconfidence reduction can be made stronger if people are required to reason why the particular option might be true. Average confidence in this experiment was more sensitive to foil plausibility (of an alternative option) and the bias (tending to underconfidence) was much lower.

The studies reviewed above were published in *Organizational Behavior and Human Decision Processes* and typically followed the experimental protocol in psychology. We will next review studies from business, economics, and finance and evaluate the presence of the problems identified in psychology and economics that are summarized in Figure 3.1.

Mahajan (1992) investigated the effects of evaluative feedback, counterfactual reasoning, and expertise in the context of marketing predictions. In the first experiment, the student subjects had to answer two alternative questions on the business material provided to them a week before the experiment. Students were also asked to state their confidence (half-range) in the correctness of each answer. After 45 days, participants were given false feedback (evaluative feedback – better or worse than average performance) on their predictions. Consequently, they were asked to repeat the task with different 70 questions. Half of the students participated in the contradictory condition – before stating their confidence students were asked to defend why the answer might be wrong. The results exhibited clear overconfidence. The unfavorable feedback resulted in insignificantly lower confidence. The contradictory evidence did not significantly decrease confidence; however it significantly increased the accuracy and consequently also decreased overconfidence. In the second experiment, the effect of expertise was evaluated. No feedback was provided, but subjects were asked to evaluate the difficulty of each question. The results suggested that people are more overconfident in tasks where they claim some expertise. The participants in the first experiment did not receive any monetary incentives, while the top performing one third of participants received $10 in the second experiment. *The results in Mahajan (1992) might be influenced by the lack of incentives for accuracy of confidence judgments and insufficient incentives for performance. The improvement in performance when asked for contradictory evidence, to some extent, supports this claim. The authors asked for confidence responses in probabilities and not for relative frequencies. Moreover, there is no theoretical model or explanation for how people should respond to false feedback, never mind what the effect of deception was.*

Wallsten, Budescu, and Zwick (1993) investigated the differences in verbal and numerical probability responses under various payoff conditions. The authors used 300 general-knowledge questions, in the form of true-or-false statements each. For each statement, the student participants had to state numerical probability (in %) that the statement was true. On another day their confidence in the same statement was expressed verbally; eleven words were selected to indicate verbal judgements (impossible - … - tossup - … - certain). The experiment consisted of four experimental sessions over four days. The authors used three payoff conditions – only flat fee, additional $ gain for best four participants ($8, $6, $4, $2, respectively), and loss of $ for worse four participants ($8, $6, $4, $2, respectively). The results revealed that the two response modes (verbal and numerical probability response mode) differ; however, one cannot say that one of them is better than the other. Numerical judgements of confidence were more unevenly spread than verbal judgements and the response distributions were more unequal in the gain than in the loss treatment. Moreover, subjects in the gain treatment exhibited higher overconfidence in their judgements than people in the other two treatments. Overconfidence was significantly higher in the verbal mode. *Wallsten et al. (1993) incorporated into their experiments all the issues that might influence the results of calibration (representative stimuli, various payoff types). The only problem might be the ambiguity of the verbal scale (across subjects), even though the authors devoted considerable effort to making the scale as identical as possible for everyone. The result (more overconfidence in verbal mode) is interesting because one might conjecture that* *people would be more comfortable with the verbal response mode. It would be interesting to see in which of the two modes calibration would improve faster if feedback were to be provided.*

Frankenberger and Albaum (1997) conducted an experiment on memory for advertisements. The authors used advertisements with three levels of involvement and two product categories (a convenience and a shopping good). The advertisement was inserted in a booklet that also contained two other ads, a two-page article on MBA degree, and a one-page article on hacky sacks. The authors manipulated the level of involvement with their instructions. Low-involvement instructions asked participants to read the MBA article within which the target advertisement was placed. Moderate-involvement instructions required participants to proofread the article and the ad and circle typographical and grammatical errors, while high-involvement instructions asked participants to imagine that they are interested in buying the advertised product and to be prepared to write a brief statement after inspecting the advertisement. After finishing the involvement tasks participants were asked to determine which of the 16 ad characteristics were actually in the target advertisement and were asked other related questions. Moreover, subjects were asked to state their confidence (full-range) in the correctness of each chosen alternative. Student participants received extra credit towards their course for participation. The results exhibited overconfidence for low-involvement for both goods and underconfidence for high-involvement only for the shopping good. In addition, the accuracy of responses increased with increasing confidence. *Frankenberger and Albaum (1997) lack incentives in their experiment. This could explain the switch from overconfidence in low-involvement treatment to underconfidence in high-involvement treatment. In addition, this experiment lacks transparent (random) choice of material and questions.*

Hoelzl and Rustichini (2005) defined overconfidence in a general way – when a majority of people estimate their skills or abilities to be better than the median. The authors used hard and easy tasks with and without monetary incentives in their experiment, for a 2x2 design. The task was to complete a gap-filling exercise of 20 sentences in a test of knowledge of vocabulary (LEWITE test). In each task participants had to choose two words from 7-9 alternatives. However, the authors did not measure confidence directly; they let a group of people vote by majority rule for one of two payoff conditions (performance or lottery). In the performance condition only participants whose performance was in the upper half of the results of all participants would be paid. In the lottery condition everyone had a 50% chance of winning money. In addition, subjects were asked for estimates of their own and group average performance. Participants (mostly students) in the monetary condition could have won money. The average vote was 55% in favor of the performance condition. From the data it followed that increasing task difficulty as well as not offering monetary incentives encouraged voting for lottery. The authors found that the behavior significantly changed with task difficulty only if money was offered (overconfidence if an easy, familiar task; underconfidence if a non-familiar task). *The basic problem of Hoelzl and Rustichini (2005) is their strategy for recruiting students, which might have caused a sample selection problem. Subjects for this experiment were recruited in classrooms and the local cafeteria until a sufficiently* *big group was gathered. Note that this is the case where relative self-assessment matters and therefore also the subject pool plays a role. This fact magnifies the improper recruitment strategy.*

Kovalchik, Camerer, Grether, Plott, and Allman (2005) examined confidence of two different subject groups (elderly individuals and students). Both groups had to state their confidence (half-range) in the correctness of each of the 20 two-alternative general-knowledge questions. The results suggested that the elderly participants were overconfident only on higher confidence levels but students were overconfident on all confidence levels. In addition, the elderly participants had a majority of confidence responses on 100 or 50 % confidence levels, which suggested higher resolution of the "experienced" group (they did better than students on the test – 74 vs. 66 % of answers were correct). *Kovalchik et al. (2005) do not report if the sample of questions was chosen randomly or not. Another problem might be the lack of incentives: students might need more incentives to think about the right estimate while older participants might have had sufficient experience to make more accurate estimates even without incentives.*

Kogan (2006) investigated the impact of overconfidence on asset markets. However, the author used a different approach than was (and is) common up to that point – he manipulated people to be overconfident and then compared their result with non-overconfident people. Student participants in the experiment were ranked according to a dice toss (non-overconfident) and according to 20 multiple-choice SAT test questions. For the latter, previous research has allegedly demonstrated that people rank themselves better than their standing warrants. Kogan (2006) did not inform his participants about their rank. The author provided participants with a signal, as follows: Participants ranked in the top half received a perfect signal, and participants ranked in the bottom half received an imperfect (noisy) signal about the value around which the signals were distributed. Then they had to make an estimate of that value. At the beginning of every round (total 20) subjects were paired (with one opponent) and new target value and signals were generated. After receiving a signal, subjects simultaneously made estimates of the target value after which estimates of all participants were publicly announced; this happened four times with the same opponent and same signals. Participants were paid according to the accuracy of their estimates. The results showed that the analogs of volume and price errors are greater in the overconfidence treatment. Based on the experimental results, the authors concluded that participants mostly strategically responde to the overconfidence of others and partly behaved overconfidently. *Kogan (2006) used overconfidence in SAT test questions as an explanatory variable for calibration in asset markets. However, as it was shown in Klayman et al. (1999), overconfidence varies across domains. Therefore the results of this experiment could be questionable (because overconfidence has been measured in a different domain). In addition, the question of the subject pool should be taken into account – it is difficult to assess one's own rank within a group of (probably) unknown students.*

To sum up, several researchers have pointed out the positive effect of feedback on calibration (e.g., Arkes, Christensen, Lai, and Blumer, 1987) and in some cases even on performance (Subbotin, 1996). Moreover, it is possible to achieve an improvement in calibration using counterfactual reasoning (Mahajan, 1992). As we already mentioned in our review of the psychology literature, random error plays a significant role too and has to be taken into account. It has also been shown that frequency responses return more accurate confidence estimates than probability responses (Price, 1998). In addition, Juslin et al. (2003) showed that there are differences in overconfidence in performance on general-knowledge tasks and sensory discrimination tasks (in this case underconfidence). We also presented studies that measure people's confidence in their performance indirectly – using the choice between a lottery and performance-based reward (Hoelzl and Rustichini, 2005). We stressed that in this kind of experiment relative self-assessment, in addition to absolute self-assessment, plays an important role. Moreover, experiments with monetary incentives deliver different results than those without financial motivation (Hoelzl and Rustichini, 2005). Wallsten et al. (1993) showed that verbal and numerical statements of confidence differ. Finally, it seems that older subjects are less overconfident than younger subjects (Kovalchik et al., 2005), which may be related to experience (either with the tasks or with confidence statement).

Despite the clear message from psychology, researchers in economics underestimate the importance of representative sampling of general-knowledge questions (e.g., Juslin et al., 2000). Even though the results by Juslin et al. (2000) suggest that representative stimuli expurgate overconfidence and the hard-easy effect for two alternative general-knowledge questions, that does not necessarily mean that overconfidence is not present if other than standard procedures and measures from psychology are used. For example, using the choice between a random lottery and performance-based payment involves the estimation of one's confidence in the correctness of one's answers and of answers of other participants in the group. This is closely tied to the unskilled-unaware problem and relative self-assessment in general. In these experiments, the question of subject pool plays an important role (see e.g., Chapter 1 of this dissertation). Therefore, researchers should be aware of the (un)representative nature of their stimuli (including representative subject sample) used in their experiments and the amount of information (especially information that subjects have about the subject pool). In line with results from psychology, it has been shown that feedback and counter-reasoning improve calibration (e.g., Mahajan, 1992). This result suggests the possible role of information availability and/or familiarity with task/elicitation method.

3.4.2 Confidence intervals

Confidence intervals are an alternative measure of people's confidence. Since confidence intervals are ranges, measuring confidence using confidence intervals requires a numerical answer (range). Thus, confidence intervals cannot be used to measure confidence in two-alternative general questions and are used to measure confidence in different types of questions. Therefore, we consider confidence intervals to be a separate paradigm. One can find the use of confidence intervals mostly in literature involving general-knowledge questions (open-ended) and forecasting tasks. Usually, researchers ask participants to estimate the 90%-confidence interval for the estimated variable. This means that the true realization of the variable should fall within the given range 90% of the time. If people set a narrower confidence interval than the true 90%-confidence interval (if the true 90%-confidence interval is measurable) or when the number of true realizations of the estimated variable falls outside the range more often than in 10% of all cases, people are identified as overconfident. Other than 90%-confidence intervals are also being used. We first review studies using open ended general-knowledge questions or similar stimuli that employ confidence intervals as a measure of confidence. Then we summarize the basic setup and findings of forecasting studies, where participants have to forecast a range for the forecasted variable so that they are *x%* confident that the forecasted variable falls within the range.

We briefly presented the results from Klayman, Soll, Gonzales-Vallejo, and Barlas (1999) in the general-knowledge questions section in psychology. In one of their experiments involving general-knowledge questions, the authors also asked student participants to state 90% confidence ranges for each answer, instead of stating subjects' confidence in the correctness of the particular answer. In this experiment, participants were paid a flat participation fee. The results suggested much greater overconfidence than in the case of direct confidence of the correctness of general-knowledge questions in this study (reported in Section 3.4.2). Similar to the case of direct confidence estimates, overconfidence differed between domains and individuals but did not linearly depend on the domain difficulty. *Klayman et al. (1999) dealt with the representativeness issue very well; however, the lack of incentives to give accurate estimates might have influenced the results.*

We also already reviewed one part of Juslin, Winman, and Olsson (2003) where the full-range confidence estimates were used. In addition, the authors also investigated participants' calibration in confidence intervals response mode. They asked participants to give an answer to the general-knowledge or sensory discrimination questions and to provide the smallest interval for which they were 50%, 90%, and 100% sure that it would be the right answer. The authors paid the participants a fixed payment. The results exhibited significant overconfidence (when measured using confidence intervals); more overconfidence was identified in the general-knowledge questions than in sensory discrimination questions. Compared to the full-range response format (2-alternative questions), the underconfidence in the case of sensory discrimination tasks changed to strong overconfidence when confidence intervals were used. Similarly, overconfidence resulting from confidence interval estimates in the general-knowledge task was several times greater than in full-range response mode. *It is not clear from the paper whether Juslin et al. (2003) used additional incentives (above flat incentives) in case of confidence interval measure as well or only in the case of full-range response measure.*

Cesarini, Sandewall, and Johannesson (2005) investigated the impact of several manipulations on the results of confidence interval estimation tasks. The authors conducted an experiment with undergraduate students who had, for each of 10 numerical questions, to state 90% confidence intervals. These authors, unlike the overwhelming majority (except for the authors of studies involving general-knowledge questions), used random sampling of questions from all possible questions in the particular domain and informed the participants about it. After answering all 10 questions, participants were asked to estimate how many of their interval estimates contained the real value and also to estimate the performance of their peers. Furthermore, participants were instructed to judge again their count of correct confidence intervals and were asked for new estimates of their and their peers' performance (number of correct confidence intervals). Subsequently, they were asked to adjust their intervals in order to make sure that they would really constitute 90% confidence intervals. In addition to the show-up fee, the authors motivated participants in the incentives treatment with extra money for each accurate estimate. The results of Cesarini et al. (2005) revealed overconfidence in both interval and frequency estimates, significantly higher in the first. Participants also anticipated overconfidence of others and iteration of frequency estimates significantly decreased the overconfidence; note that this happened without any feedback. Moreover, the results did not suggest a significant effect of monetary incentives (besides lowering the above average effect). *Cesarini et al. (2005) showed that overconfidence in confidence intervals can be reduced through repeatedly requesting the answer (and thus forcing participants to think through their answers over and over again). High initial overconfidence (which is significantly decreased without providing any additional information or feedback) might be the consequences of people not being accustomed to this response format.*

Some researchers introduced the confidence intervals measure of overconfidence also in the experimental asset market environment. For example, Kirchler and Maciejovsky (2002) conducted an experiment where 72 student participants traded one risky asset in six markets. The authors manipulated dividend information (complete and no information) and public signals about the market participants' average price prediction (precise, vague, no signal). First, the authors measured participants' risk aversion. In the main experiment, participants made bids and asks in an anonymous double auction to trade the risky assets. Before the market was opened, participants received either information about dividends distribution in the next round or no information and were asked to predict the average market price, state a 98% confidence interval, and state the confidence in their prediction on a 9-step scale. Consequently they received one of three public signals and then the market was opened. Participants were paid according to their trading performance in the experiment. The results suggested learning, because the observed average market prices came closer to expected prices with increasing experience (over time). The number of concluded contracts declined, the number of not accepted offers increased, and prices decreased over time. Surprisingly, Kirchler and Maciejovsky (2002) identified slight risk aversion which was not significant among sessions. The results revealed information diffusion in the experiment. The participants in this experiment were highly overconfident in creating their confidence intervals (except for the first trading period) and their overconfidence increased over time. However, participants were not generally overconfident when overconfidence was measured on the 9-step scale (some were underconfident and some well-calibrated during some periods). Kirchler and Maciejovsky (2002) also found that trading volume was positively correlated with individual earnings and that subjects identified as overconfident in creating subjective confidence intervals earned significantly less than others. *The only point to question in the experiments reported in Kirchler and Maciejovsky (2002) is whether the use of student participants is appropriate. They typically do not have* *experience with market and especially with the creation of confidence intervals in this environment (note that about one fifth of them were social science students). Thus, this fact could have contributed to the observed bias.*

Biais, Hilton, Mazurier, and Pouget (2005) experimentally measured the degree of overconfidence in judgment and self-monitoring. Student participants traded a single asset with uncertain dividend paid at the end of the trading game with asymmetric information and with permitted short-selling. They received heterogeneous private signals before trading. The asset was traded in oral double auction and call auction. The authors measured overconfidence outside the asset market: They asked participants to state a 90% confidence interval for each of 10 questions. The authors then used the overconfidence measure – the proportion of answers falling outside the stated range – in the regression analysis as an explanatory variable for performance and trading behavior. Participants were told that their final grade would also reflect their results in the experiment. Biais et al. (2005) found that their participants were greatly overconfident (only 64% of all confidence intervals contained the true value). Moreover, the results suggested that miscalibration (overconfidence) reduced and self-monitoring (the disposition to attend to social cues) enhanced trading performance. *The experiment in Biais et al. (2005) has several shortcomings. First and most important, overconfidence was measured artificially outside the market. This is not a very fortunate choice because e.g., Klayman et al. (1999) showed that overconfidence differs over domains. Second, the incentives might not have been sufficient. Third, the authors did not control for risk aversion which might play role in experiments involving uncertainty. Fourth, similar to the previously reviewed paper, the use of student participants may be problematic.*

Oenkal, Yates, Simga-Mugan, and Oetzin (2003) inspected overconfidence in predictions of foreign exchange. The professionals and students had to make point forecast, directional forecast (and to state half-range confidence in their forecast) and to state 90% confidence intervals for one-day and one-week-ahead foreign exchange forecasts. Participants volunteered and did not receive any financial incentives. In the point estimation tasks, professionals performed significantly better than amateurs. The estimates of performance of both groups were about 20% but the actual hit rate was 6% for professionals (and 2% for amateurs), suggesting strong overconfidence. Directional predictions were more accurate in the one-week-ahead forecasts while point estimates were more accurate in one-day-ahead forecasts. Participants were identified as overconfident in the case-level directional forecasts. However, they turned out to be underconfident in aggregate level, when overconfidence was measured as the difference between expected and realized percentages. The overconfidence in confidence intervals response mode (too-narrow confidence intervals) was much higher than in point and directional estimation of foreign exchange. Subjects were more accurate in creating confidence intervals for one-day-ahead forecasts. In addition, subjects were asked to estimate how many of the fifty 90%-onfidence intervals would contain the true value. If the overconfidence was measured as the difference between this estimate and the actual hit rate, it was significant only for amateurs making one-week-ahead forecasts. *Oenkal et al. (2003) did not use incentives to motivate their subjects for more accurate confidence intervals/predictions. This claim is supported by the finding that people, after stating*

90%-confidence intervals, estimate their "hit rate" only around 60%. This low confidence, however, might also be due to non-familiarity with confidence intervals measure.

Bolger and Oenkal-Atay (2004) explored the effect of calibration and environmental feedback on overconfidence. In their experiment, business students had to state confidence intervals for one-period-ahead forecasts of altogether 96 time series. The series differed in various characteristics. During the first session, participants were asked for one-period-ahead forecasts in the form of confidence intervals and they could choose a confidence percentage for each session. In the second and third session (each after three days), they received feedback on the performance of their estimates and made estimates of another 32 time series. Finally after another 3 days they received feedback and completed a questionnaire. Participants performing in this experiment received extra credit in a forecasting course. The results suggested strong overconfidence that was significantly decreased (in some cases even slightly reversed) with feedback. The authors also identified a significant effect of feedback on improvement in the ability to match probabilistic responses to changes in uncertainty in the time series. In addition, participants increased confidence interval width and percentage confidence slightly over time. *Insufficient incentives could have caused participants in the experiments to exert insufficient effort in stating accurate confidence intervals. It is also not clear what kind of representation of information is more suitable: graphs (as these authors did,) basic statistics lists of the time series, or both.*

Du and Budescu (2007) asked participants in their experiments for point estimates and interval forecasts of future values of low- and high-volatility time series. The authors provided their subjects with stock prices of a year period and asked them to make forecasts at the end of Month 3 of the subsequent year. In the case of confidence intervals, participants had to state 50%, 70%, and 90% confidence intervals. At the end of the experiment, participants took part in a lottery where the more accurate ones had a higher chance to win money. The results revealed that the low-volatility time series predictions were more accurate than the high-volatility time series. The authors identified underconfidence in 50%, good calibration in 70%, and overconfidence in 90% confidence interval predictions. Moreover, the reported confidence intervals were skewed. *Du and Budescu (2007) used selected time series and it is not clear how representative this selection was. The authors displayed the information in the form of graphs – it again raises the question of representative information format.*

Similarly, Budescu and Du (2007) investigated calibration and the consistency of direct probability judgments and confidence interval estimates. 63 graduate students were provided with graphs of twelve monthly prices of Year 1 and asked to make a series of judgments about future prices (Month 3 of Year 2). Specifically, they were asked to estimate the probability that the future price will exceed $20, the lower and upper bounds (i.e. an interval estimate) of this probability, 50%, 70%, and 90% confidence intervals, and the median price. Participants participated in lotteries for $50, where the more accurate predictions implied higher probability of winning the prize. Four participants won the prize. The results revealed that participants were overconfident in the case of direct probability estimates and the 90% confidence intervals, but well-calibrated at the 70% and underconfident at the 50% confidence interval. The authors noted that the difference in calibration in confidence intervals is due to the trade-off between accuracy and informativeness – wider intervals are less informative, but are likely to be more accurate (see also Yaniv and Foster, 1995, 1997). *Budescu and Du (2007) used graphical presentation of the price history; presenting numerical values as well might be a more natural form of information for people.*

Meloy, Russo, and Miller (2006) were primarily interested in the impact of monetary incentives on performance. In one of their experiments, the authors investigated the effect of incentives on overconfidence. They asked student participants to state the upper and lower limit of the interval about which they were 90% confident that the right answer to a general-knowledge question falls inside the interval. Students were asked to state 90%-confidence intervals to 10 questions; the more intervals that were stated correctly, the more money they (in the incentives treatment) received, yet if all were correct, they received only about half of the money for 9 correct intervals. Participants in the non-incentives treatment were paid a flat participation fee. The results revealed that participants in the incentives treatment spent more time on these questions; however, overconfidence was much higher than in the non-incentives treatment. The authors concluded that incentives had elevated participants' mood and this increased overconfidence. This effect was slightly mitigated by the increase in effort. *Meloy et al. (2006) do not say how they chose the questions (randomly or not). A selective choice of the questions could have shifted the results in favor of overconfidence.*

The results from these studies on confidence intervals suggest much stronger overconfidence than conventional measures of confidence. Klayman et al. (1999) showed that overconfidence is not linearly related to task difficulty yet it differs across domains. This was confirmed by Juslin et al. (2003), who found much greater overconfidence in general-knowledge questions than in sensory discrimination tasks. It was also shown (Du and Budescu, 2007) that overconfidence is present mainly in high-confidence confidence intervals (e.g., 90%) while underconfidence prevails in low-confidence confidence intervals (e.g., 50%). The effect of feedback (Bolger and Oenkal-Atay, 2004) and even pure repetition of confidence intervals estimation (Cesarini et al., 2006) on calibration was found to be significantly positive also in experiments involving confidence intervals. Moreover, people were identified as overconfident in various prediction tasks in financial markets (e.g., Kirchler and Maciejovsky, 2002). Also in forecasting tasks, confidence intervals produce greater overconfidence than confidence in point estimates (e.g., Oenkal et al., 2003) and feedback lowers overconfidence. Lastly, experts seem able to create more accurate confidence intervals than non-professionals (Oenkal et al., 2003).

Several studies showed that an overconfidence measure with confidence intervals varies across domains and that it is not a universal bias (e.g., Klayman et al., 1999). It therefore seems inappropriate to measure overconfidence outside the task in focus as Biais et al. (2005) did. But why are people less calibrated when confidence intervals are used? One reason might be that people are not accustomed to the confidence interval notion – they do not use confidence intervals in ordinary life. In a real life situation, people more often express their confidence in the form of a statement or a choice than they do in the form of a confidence interval. The above-mentioned positive impact of feedback on calibration supports this claim. In addition, Moore and Healy (2008) found that overprecision (too-narrow confidence intervals) was correlated with overestimation and that overprecision significantly decreased with information and experience (from 55% to 85% in the case of 90%-confidence intervals). Moreover, confidence in answers to two-alternative general-knowledge questions is a linear measure on a range 50-100% (from guessing to sure) while confidence intervals are in general not linear (there is much greater probability of the event in the range around the mean than further away from the mean where it decreases non-linearly). Therefore, it might be more difficult to account for this non-linearity and to handle it correctly. Moreover, it was shown (e.g., Du and Budescu, 2007) that people are not able to discriminate between the precisions of the confidence intervals – they are well-calibrated when stating 70% but less calibrated when stating 90% and 50% confidence intervals (their intervals are closer to the 70% than they should be). Yaniv and Foster (1995, 1997) suggested that interval forecasts are a trade-off between accuracy and informativeness. The idea is that people prefer informativeness to accuracy (i.e. giving more informative narrower intervals rather than non-informative wide intervals with desired accuracy). Recently, Juslin, Winman, and Hansson (2007) introduced the notion of a naive intuitive statistician and developed the naive sampling model, which assumes that although people accurately acquire the sample information, they take sample properties as estimates of population properties. The authors showed that this model is able to explain main patterns of the experiments on confidence in half-range and full-range mode as well as in confidence interval measures. Juslin et al. (2007) also devised a method for minimizing the overconfidence in the produced intervals – to create confidence intervals for unknown quantities where participants rely on proportion rather than coverage as the estimator variable.

3.4.3 Forecasting

As already mentioned at the beginning of the previous section, overconfidence was identified in studies involving forecasting of various events. These studies use various forecasting stimuli but the common feature is, similar to general-knowledge tasks, that people have to forecast a state of an event and state their confidence in the correctness of their forecast. Forecasting of future events is widely used in many professions (e.g., economics, finance, weather forecasting, betting). Since the future events usually depend on some stochastic process, it is almost impossible to state an accurate point estimate (unless there is some finite range of possibilities). There are two ways to measure confidence in the forecast: confidence in the point forecast (or directional forecast) and confidence interval estimate. In this section, we will review studies of both groups, even though we reviewed studies involving the confidence interval measure of confidence separately (along with non-forecasting studies) in the previous section.

Bolger and Harvey (1995) investigated the calibration of students in determining whether the value of a predicted variable will be below or above a given value. The authors constructed time series with and without trend and graphically presented the first 31 points. They confronted participants with 7 different values corresponding to different probabilities, one after another. For each value, the authors asked their participants to state the probability that the real value would be below (above – in the second experiment) that value. The results revealed overestimation of probabilities lower than 0.5 and underestimation of probabilities greater than 0.5. Over- and underestimation was greater for trended than non-trended time series and also greater for above- than for below-judgments. The authors concluded that subjects in these experiments exhibited underconfidence in their estimates of where the next point would lie. *Bolger and Harvey (1995) do not report use of any incentives in their experiments; this might have increased the inaccuracy of subjects' responses. In addition, the authors ask for probability estimates; use of frequencies could improve calibration.*

Bolger and Oenkal-Atay (2004) explored the effect of calibration and environmental feedback on overconfidence. In their experiment, business students had to state confidence intervals for one-period-ahead forecasts of altogether 96 time series. The series differed in various characteristics. During the first session, participants were asked for one-period-ahead forecasts in the form of confidence intervals and they could choose a confidence percentage for each session. In the second and third session (each after three days), they received feedback on the performance of their estimates and made estimates of another 32 time series. Finally after another 3 days they received feedback and completed a questionnaire. Participants performing in this experiment received extra credit in a forecasting course. The results suggested strong overconfidence that was significantly decreased (in some cases even slightly reversed) with feedback. The authors also identified a significant effect of feedback on improvement in the ability to match probabilistic responses to changes in uncertainty in the time series. In addition, participants increased confidence interval width and slightly increased percentage confidence over time. *Insufficient incentives could have caused the participants in the experiments to exert insufficient effort in stating confidence intervals. It is also not clear what kind of representation of information is more suitable: graphs (as these authors did, basic statistics lists of the time series, or both.*

We already reviewed Kirchler and Maciejovsky (2002) above. To recall, the authors manipulated dividend information (complete and no information) and public signals about the market participants' average price prediction (precise, vague, no signal). First, the authors measured participants' risk aversion. In the main experiment, participants made bids and asks in an anonymous double auction to trade the risky assets. Before the market was opened, participants received either information about dividends distribution in the next round or no information and were asked to predict the average market price, state the 98% confidence interval, and state their confidence in the prediction on a 9-step scale. Consequently they received one of three public signals and then the market was opened. Participants were paid according to their trading performance in the experiment. The results suggested learning, because the observed average market prices came closer to expected prices with increasing experience (over time). The number of concluded contracts declined, the number of not accepted offers increased, and prices decreased over time. Surprisingly, Kirchler and Maciejovsky (2002) identified slight risk aversion, which was not significant among sessions. The results showed that information diffusion happens in the experiment. The participants in this experiment were highly overconfident in creating their confidence intervals (except for the first trading period) and their overconfidence increased over time. However, participants were generally not overconfident when overconfidence was measured on the 9-step scale (some were underconfident and some well-calibrated during some periods). Kirchler and Maciejovsky (2002) also found that trading volume was positively correlated with individual earnings and that subjects identified as overconfident in creating subjective confidence intervals earned significantly less than others. *The only question is whether the use of student participants is appropriate. They usually do not have experience with a market and especially with the creation of the confidence intervals in this environment (note that about one fifth of them were social science students). This fact could have contributed to the observed bias.*

Oenkal, Yates, Simga-Mugan, and Oetzin (2003) inspected overconfidence in predictions of foreign exchange. Professionals and students had to make point forecast, directional forecast (and to state half-range confidence in their forecast) and to state 90% confidence intervals for one-day and one-week-ahead foreign exchange forecasts. Participants volunteered and did not receive any financial incentives. In the point estimation tasks, professionals performed significantly better than amateurs. The estimates of performance of both groups were about 20% but the actual hit rate was 6% for professionals (and 2% for amateurs), suggesting strong overconfidence. Directional predictions were more accurate in the one-week-ahead forecasts while point estimates were more accurate in one-day-ahead forecasts. Participants were identified as overconfident in the case-level directional forecasts. However, they turned out to be underconfident in the aggregate level, when overconfidence was measured as the difference between expected and realized percentages. The overconfidence in confidence interval response mode (too-narrow confidence intervals) was much higher than in point and directional estimation of foreign exchange. Subjects were more accurate in creating confidence intervals for one-day-ahead forecasts. In addition, subjects were asked to estimate how many of the 50 90%-onfidence intervals would contain the true value. If the overconfidence was measured as the difference between this estimate and the actual hit rate, it was significant only for amateurs making one-week-ahead forecasts. *Oenkal et al. (2003) did not use incentives to motivate their subjects for more accurate confidence intervals/predictions. This claim is supported by the finding that people, after stating 90%-confidence intervals, estimate their "hit rate" only around 60%. This low confidence, however, might be also due to non-familiarity with the confidence interval measure.*

Thomson, Oenkal-Atay, Pollock, and Macaulay (2003) investigated experts' and students' performance in predicting simulated currency series with various positive and negative trends (mild, medium, strong, and very strong). The data was graphically presented for past 60-month periods. Participants were not financially motivated and were asked to make a directional prediction for the subsequent month and state their confidence in the correctness of their answer (50%-100% where 50% meant no change). The authors identified a significant effect of trend on performance (better performance for stronger trends). The results also suggested better relative accuracy and higher profitability for students. Regarding overconfidence, experts were more overconfident on the mild trends while more underconfident on the strong and very strong trends (hard-easy effect). Participants were more confident on positive trends than on negative when trends were strong, and less confident on negative than positive trends when these were weak. *Thomson et al. (2003) did not offer any incentives in their experiment. Moreover, they allowed their participants to complete the task outside the laboratory, thus losing the advantage of a controlled laboratory environment. It is also not clear how representative the simulated currency series were.*

Tyszka and Zielonka (2002) investigated the predictions of two groups of experts – financial analysts and weather forecasters. Financial analysts had to forecast the value of the Warsaw Stock Exchange Index in one month, while weather forecasters were asked to predict the average monthly temperature (the month starting in two weeks). Both groups were given the possibility to choose one of three equally probable intervals (but they were not informed that these intervals were equiprobable). First, the participants had to state how confident they were in their choice. Then, they had to rank-order the intervals and assign probabilities to each of them (to sum up to 100%). After two months, the respondents received feedback on their performance and were asked to assess the importance of several factors that might lead to wrong answers. Subsequently, they were asked to make similar predictions as in the first period for the next period. The authors did not offer any financial incentives to participants. Only one-third of financial analysts and two-thirds of weather forecasters made correct forecasts (note that weather forecasters frequently marked more intervals as equiprobable). Mean self-evaluation in both groups was similar and both groups revealed overconfidence, which was, due to worse performance, significantly higher in the group of financial analysts. Weather forecasters attached significantly higher importance to the probabilistic argument (that there is no certainty that the event in question can be predicted accurately). After evaluation of reasons why could the forecast fail, the weather forecasters, unlike financial analysts, lowered their assessment of ability to predict the event. *Tyszka and Zielonka (2002) did not use any incentives in their experiments. Therefore, it remains a question what amount of effort the experts exerted in making their forecasts and stating their confidence in these forecasts.*

Torngren and Montgomery (2004) conducted experiments with two groups – stock market professionals and laypeople (students) who were provided with information about the stock (name, industry, monthly %-price change for the past 12 months) and asked to make one-month-ahead predictions of the rate of change for the share prices of 20 stocks. Furthermore, they were asked to predict their own, own group and other group's errors and to rate the extent to which they used each of the following four strategies: previous monthly results, other knowledge, intuition, and guessing. The authors did not provide monetary incentives. The results suggested that professionals had higher expectations about their abilities to make accurate predictions. Performance of professionals was significantly lower than of laypeople and even worse than chance. Both groups exhibited overconfidence; however, professionals were more overconfident than laypeople. While laypeople were mostly guessing, professionals were relying on expert knowledge. The errors were approximately the same in both groups (as predicted for laypeople); both groups, however, thought that the errors of professionals would be half of the laypeople's errors. *Torngren and Montgomery (2004) did not use any kind of incentives in their experiment. This might have caused the professionals not to perform to their best knowledge or to have overly high expectations/estimates of their performance.*

Andersson, Edman, and Ekman (2005) investigated the differences in performance and confidence of experts and non-experts in predicting the outcome of the first round of World Cup 2002. The authors used 2 types of questionnaires – one with cues about the 32 national teams and one with no information. Participants were asked to state how well they knew each teams, for each group of four teams they predicted which teams would qualify and stated their confidence on a seven-point verbally anchored scale. In addition, after completing the task for every group, participants were asked to predict how many of the 16 teams chosen in the previous task would actually qualify. Finally, they were asked to rate their global knowledge of soccer and ability to predict soccer matches. The authors identified four groups of participants: experts, knowledgeable Swedish students, naïve Swedish students, and American students. All groups but experts were paid a flat participation fee with extra reward for the most accurate estimates. The results showed that experts' performance was not better than the performance of the other groups. All groups did better than chance; however, the simple rule (according to the ranking) outperformed the participants of the experiment. The authors used three measures of overconfidence: expected accuracy (frequency), overconfidence, and aggregate confidence in forecasts (average confidence). All three measures revealed that experts were more confident than non-experts, resulting in overconfidence of experts. The information had two effects: increased accuracy of American students and increased the aggregated confidence of non-experts. *The experiments in Andersson et al. (2005) were actually conducted in the form of questionnaires, though as they involved experimental manipulation we included this study into the review. Even though the authors noted that experts should be driven by intrinsic motivation, the results suggest that they either did not have sufficient motivation or that their expertise did not give them an advantage over non-experts.*

Du and Budescu (2007) asked participants in their experiments for point estimates and interval forecasts of future values of low- and high-volatility time series. The authors provided their subjects with stock prices of a year period and asked them to make forecasts at the end of Month 3 of the subsequent year. In the case of confidence intervals, participants had to state 50%, 70%, and 90% confidence intervals. At the end of the experiment, participants took part in a lottery where the more accurate ones had a higher chance to win money. The results revealed that the low-volatility time series predictions were more accurate than the high-volatility time series. The authors identified underconfidence in 50%, good calibration in 70%, and overconfidence in 90% confidence interval predictions. Moreover, the reported confidence intervals were skewed. *Du and Budescu (2007) used selected time series and it is not clear how representative this selection was. The authors displayed the information in the form of graphs – it again raises the question of representative information format.*

Similarly, Budescu and Du (2007) investigated calibration and the onsistency of direct probability judgments and confidence interval estimates. 63 graduate students were provided with graphs of twelve monthly prices of Year 1 and asked to make a series of judgments about future prices (Month 3 of Year 2). Specifically, they were asked to estimate the probability that the future price will exceed $20, the lower and upper bounds (i.e. an interval estimate) of this probability, 50%, 70%, and 90% confidence intervals, and the median price. Participants participated in lotteries for $50, where the more accurate predictions implied higher probability of winning the prize. Four participants won the prize. The results revealed that participants were miscalibrated – overconfident in the case of direct probability estimates and the 90% confidence intervals, but well-calibrated at the 70% and underconfident at the 50% confidence interval. The authors noted that the difference in calibration in confidence intervals is due to the trade-off between accuracy and informativeness – wider intervals are less informative, but are likely to be more accurate (see also Yaniv and Foster, 1995, 1997). *Budescu and Du (2007) used graphical presentation of the price history; presenting numerical values as well might be a more natural form of information for people.*

Kelemen, Winningham, and Weaver (2007) investigated calibration in predictions of future recalls of Swahili-English vocabulary items and the impact of repeated study on calibration. Student participants volunteered in the experiment (received additional course credit) that took place in 5 consecutive sessions. The mean confidence decreased and accuracy increased over time with practice without any feedback. Participants mostly improved their calibration on items that they would not recall later. The authors found no difference in the confidence of people with different SAT scores but there was a difference in recall performance – resulting in higher overconfidence of students with lower SAT scores. These students also adjusted their predictions less effectively. *Kelemen et al. (2007) used course credit as incentives, which might not have been sufficient.*

To sum up, overconfidence in forecasting has been extensively investigated in the last decade and the present research suggests mostly overconfident behavior, especially if confidence intervals are forecasted (e.g., Bolger and Oenkal-Atay, 2004). However, confidence in the correctness of confidence intervals or of point estimates is not necessarily excessive (e.g., Oenkal et al., 2003). As our review shows, overconfidence usually decreases with feedback or practice (Kelemen et al., 2007). People tend to overestimate probabilities lower than 0.5 and overestimate probabilities higher than 0.5 (Bolger and Harvey, 1995); note the similarity with the hard-easy effect from general-knowledge questions. As for expertise, economic experts seem to perform on forecasting tasks worse than students, yet experts exhibit higher confidence and consequently also overconfidence (e.g., Thomson et al., 2003). However, weather forecasters exhibit much lower overconfidence than economists (Tyszka and Zielonka, 2002).

The problem in forecasting studies might be the method used to measure overconfidence. Predictions of subjects should be compared with the best possible prediction given all the available information and not with the actual result (which might be biased in one or other direction). For example, comparison of the predictions of stock prices based on some price history with actual prices might lead to overestimation if the actual prices unexpectedly decreased (due to some external shock that was unpredictable) – and thus we would observe overestimation of prices even if the predictions were created "optimally".

What is the reason that experts perform worse than students? It might be un-representativeness of the stimuli material, which is usually selected (non-randomly) by experimenters. Alternatively, the lack of (sufficient) incentives for experts might explain this surprising result. It is also striking that the performance of economists is in some experiments lower than chance. This suggests that there are some serious problems either with the experts or with the experimental design and/or implementation in experiments on forecasting (e.g., insufficient incentives). Note that the problem with calibration of experts from the economics field is not universal for all experts. For example, experienced weather forecasters are well-calibrated (e.g., Murphy and Winkler, 1984), as well as horse race bettors (e.g., Bruce and Johnson, 2001) or bridge players (e.g., Keren, 1997). Therefore, further investigation of this problem is needed.

3.4.4 Market-Entry Games

Market-entry games are a class of non-cooperative games where N players have to choose simultaneously whether or not to enter a certain market (usually with pre-announced capacity c). The players who do not enter the market receive a fixed payoff K. The payoff of players who enter the market depends on market capacity c and the number of entering players E; in the general case it is equal to K+rK(c-E). Thus, the payoff from entering the market is higher than from staying out only when number of entrants is less than capacity c. This simple setup was initially used by Kahneman (1988) who observed very surprising results – the number of entrants in this non-cooperative game was typically in the range [c-2, c+2]. Erev and Rapoport (1998) investigated the effect of altering the information feedback in market-entry games with constant market capacity and various forms of feedback – individual payoffs privately displayed, individual payoffs publicly displayed, or individual payoffs publicly displayed and payoff rule explained. The results suggested excess entry in the first period of any treatment. However, the convergence to equilibrium was observed in almost all treatments. The authors also found a significant information effect when information about the payoff of others caused an increase in number of entry. Subsequently, Rapoport, Seale, and Winter (2002) considered market-entry games with full information. The subjects assigned to one of the 5 types of market-entry cost played 100 trials with randomly chosen capacity (from 10 different values) for each trial. After each trial each subject received feedback on the total number of entrants in that trial, his own payoff, and his own cumulative payoff. The authors identified coordination on the aggregate level and obvious learning over time. They observed differences among types of players as well as among people of the same type. The omission of information about asymmetry (about the differential entry fees) caused an improvement of coordination between types while the aggregate coordination stayed unchanged.

Duffy and Hopkins (2005) were concerned with the question what kind of equilibria people play in market-entry games with equal market-entry cost. The authors conducted three treatments: one's own payoff without knowing the payoff function, one's own payoff and the payoff function, and aggregate information plus actions of all other players. The authors identified one pure strategy and one symmetric mixed strategy Nash equilibrium in this market-entry game. The results showed that the more information provided, the earlier the pure strategy Nash equilibrium was achieved. Furthermore, while information seemed to have an effect on the convergence speed, it did not affect the equilibrium choice. In general, then, the basic research on market-entry games showed quick convergence to some of the equilibria, yet without being able to distinguish whether individuals were playing a pure strategy or symmetrical mixed-strategy equilibria. Several researchers extended the major research line in various dimensions (e.g., impact of feedback or information on the market-entry decisions).

Camerer and Lovallo (1999) introduced a new feature into market-entry games –relative abilities of players. The authors experimentally tested the hypothesis that frequently observed business failures are a result of managers acting on the optimism about their relative skills. In the experiment, subjects had to choose whether or not to enter a market with a pre-announced capacity. The authors extended the common market-entry games in four ways: rank-dependent payoffs; rank dependent on subject's skill or chance device; controlling for self-selection; and forecasting the number of entrants. The most important feature was the skill-dependent payoffs, which captured the differences in managerial skills. The payoff structure differed from the usual structure as follows: the number of best ranked entrants up to the capacity value shared $50 proportionally, with higher ranked entrants earning more; the remaining entrants lost $10. Subjects were initially endowed with $10, thus non-entering subjects kept $10. Every subject was ranked according to random device result and according to his/her performance on a skill or trivia task. The student subjects played 12 rounds of each (random ranking and skill-dependent ranking) condition. In each round, subjects also forecasted how many entrants they expected to enter, and made the decision to enter or not. Then they were informed how many entrants entered in that round (no further feedback was provided). After the 24 rounds, subjects solved puzzles or took the trivia quiz and random ranking was determined. Students were paid according to their performance in the experiment. The results revealed that when people bet on their own relative skills there was more entry and lower industry profit than in the random ranking treatment. Moreover, the reference group neglect caused larger skill-random entry differences with self-selected subjects (self-selection makes the overconfidence effect stronger). Camerer and Lovallo (1999) also showed that the excessive entry was not caused by underestimation of the number of entrants. The authors concluded that overconfidence about one's own relative skills is the main cause of excessive entry into the market. *Camerer and Lovallo (1999) did not control for risk aversion, which might matter in this setting because starting a business is a risky choice in an environment with uncertainty (unknown skills of others). In addition, a fixed number of entrants is more common for public tenders (with capacity 1) while in markets there often is a given profit that can be shared among any number of entering firms (where the better ones earn more and the worse ones earn less or even lose money). Therefore one could question the external validity of this experiment.*

Hogarth and Grieco (2004) chose a different approach to explain excess entry. The authors claimed that entry decisions are ambiguous gambles and that overconfidence plays no role. To support these claims, they conducted an experiment where student subjects first answered 20 general-knowledge and logic questions of two levels of difficulty: easy (choose between 2 alternatives) and difficult (choose among 5 alternatives). Consequently, the subjects were asked to estimate how many questions they answered correctly. Then they faced six choices between non-ambiguous and ambiguous lotteries. After participants received feedback on their performance (number of correct answers) on general-knowledge and logic questions, they faced the same choices in random order again. The gambles consisted of chance of winning or losing money; 50%-50% in non-ambiguous gambles and with unknown probabilities in ambiguous gambles (where half of participants were told that the probability of winning money equals their percentile ranking in the general-knowledge and logic questions quiz). The lotteries were designed to capture subjects' attitudes toward ambiguity. Participants were paid according to their choices and gamble results. The results suggested that when ambiguous probabilities are related to domains known to subjects, then subjects make more ambiguous choices and that overconfidence as well as task difficulty plays no role. In addition, Hogarth and Grieco (2004) found support for reference group neglect (after receiving feedback, participants decreased ambiguity-seeking in the harder tasks yet not in the easier tasks treatment) identified already by Camerer and Lovallo (1999). The authors also suggested that people might gain utility from participating in activities in which they feel competent. *Hogarth and Grieco (2004) did not control for risk aversion, which might matter in these situations. In addition, it is not clear if the questions were chosen randomly or not.*

Moore and Cain (2007) suggested that Bayesian explanation can account for the observation that people believe to be below average on skill-based tasks that are difficult. The authors claimed that people have better information about themselves than about others and therefore have less extreme beliefs about others than about their own performance. In the experiment, student subjects were ranked according to a simple or difficult quiz or by random device. They were paid according to their performance in the market-entry games. Each round, participants were first ranked, then made their entry decision, answered questions regarding their own and others' performance, and finally received full feedback. The results revealed that people entered markets most frequently in the simple markets, less frequently in random ranking, and the lowest entry rates were observed in the difficult markets. In a simple market, people underestimated their score and expected more competition, yet despite this they entered. In difficult markets, people overestimated the performance of others, which deterred their entry in spite of the correct estimation of a low number of entrants. These results were consistent throughout the experiment even when subjects received full feedback. The analysis also showed that participants weighted more beliefs about their own scores than about others' scores, which supported the initial hypothesis. *Moore and Cain (2007) did not control for risk aversion of their participants. Moreover, the authors neither sampled the questions randomly nor stated the reference class. The use of student subjects questions the external validity of these experiments because the skill distribution of this group might be very different from those in real markets.*

Moore, Oesch, and Zietsma (2007) investigated the behavior of actual entrepreneurs (founders), working professionals which considered starting a business (non-founders), and students in market-entry games. First, the authors reported the results of a survey in which they identified the factors decision makers take into account when making the venturing decision. The authors determined that founders as well as non-founders based their decisions on their own evaluation and neglected external factors. Second, Moore et al. (2007) conducted an experiment where they manipulated task difficulty and market capacity. Eight people played four rounds of a market-entry game where the entry decision depended on the result of a trivia quiz (only top-ranked entrants earned positive profit). Only students participated in this experiment and were paid according to their performance and decisions. The only feedback they received was the amount of money won after 2 rounds. The results showed that people entered the market more often in simple tasks than in harder tasks treatments (69% vs. 39%) and independently of the market capacity. Excess entry (insufficient entry) in easy (difficult) task treatment happened despite the fact that participants correctly anticipated excessive (insufficient) entry of others. *Moore et al. (2007) did not control for risk aversion. Again, the use of students might reduce the ecological validity of the experiment.*

The research in market-entry games has undergone two phases. First, the impact of feedback and information was investigated in repeated market-entry games. It was shown that people are well-calibrated in their market-entry decisions and that with more information, convergence is achieved faster (e.g., Erev and Rapoport, 1998). Second, after introducing skill-dependent payoff (Camerer and Lovallo, 1999), the effect of task difficulty was studied alongside feedback and information. The research suggests that people are overconfident in their own skills in easy tasks (markets) – resulting in excessive entry and underconfident in difficult tasks (markets) – resulting in insufficient entry (e.g., Moore et al., 2007). The results also suggest that feedback might not always correct this biased behavior (Moore and Cain, 2007).

Our main criticism of the research on market-entry games is that the researchers do not pay attention to issues from psychology, namely representativeness of stimuli (or at least they do not report how the questions were selected; e.g., Hogarth and Grieco, 2004). To determine the relative standings these authors mostly use the well-investigated general-knowledge questions. However, they use a low number of questions that were not chosen by random sampling from the particular environment (e.g., 10 questions in Camerer and Lovallo, 1999; 6 questions in Moore and Cain, 2003). If one combines the low number of questions to determine the ranking with the error that people make in their judgments (e.g, Erev et al., 1994; Soll, 1996) then the relative ranking is an unreliable measure of people's beliefs. However, most of the results show, similarly as in the market-entry games without skill-dependent payoff, that with sufficient feedback and information, overconfidence significantly decreased or even diminished quite quickly; even though feedback was not always an efficient tool in decreasing overconfidence in all experiments. Note that the dependence of over- and underconfidence on task difficulty is at odds with Juslin et al. (2000), who showed that task difficulty does not influence overconfidence in absolute self-assessment on 2-alternative general-knowledge questions.

Moreover, risk aversion is likely to play an important role in market-entry games as these represent a risky situation involving uncertainty (about the absolute/relative abilities of others). Therefore, researchers should include risk aversion measures into this class of experiments.

In addition, we will briefly review a study that uncovers possible reasons for the appearance of overconfidence and reference group neglect in market-entry game experiments. Elston, Harrison, and Rustroem (2005) conducted two skill treatments (simple and difficult general-knowledge questions) with market capacity equal to one. After making the entry decision, subjects completed general-knowledge tasks, estimated the number of entrants, and their own ranking in the task. The results suggested that task difficulty and risk attitudes have no effect on the probability of entry. However, the authors found a significant effect of the predicted number of entrants and the subject's confidence in their skill. In addition, part-time entrepreneurs were identified as being 35% more likely to enter (after controlling for risk attitudes) than full-time entrepreneurs or non-entrepreneurs. Besides, Elston et al. (2005) found that entrepreneurs are less risk averse than non-entrepreneurs and exhibit a significant joy of winning. The results suggest two possible explanations of reference group neglect: either it might be a laboratory (experimental) artifact, or it might apply only to a selected group of people (so called part-time entrepreneurs in Elston et al., 2005), or both. In addition, note that Elston et al. (2005) used market capacity equal to one – only the subjects with the best ranking should enter. Recall the results of Kruger and Dunning (1999) that suggest underestimation of the best performing (skilled) participants' ranking. These results (underestimation) are clearly in contradiction with Elston et al. (2005), who identify good calibration (for entrepreneurs and non-entrepreneurs) and overconfidence (for part-time entrepreneurs). The most important question remains: How representative are the reference groups (students) in a laboratory setting?

3.4.5 Auctions

Overconfidence has also been identified in people's behavior in finance. Several empirical studies have identified investors as overconfident (e.g., Barber and Odean, 2000, 2001; Menkhoff, Schmidt, and Brozynski, 2006). Some researchers, encouraged by these empirical findings, decided to investigate the behavior in experimental asset markets. These markets allow them to inspect people's trading behavior under various conditions and information structure. Auctions are usually used to trade assets in experimental asset markets. In this section we review a few experiments in which participants trade assets in the laboratory setting.

We already reviewed Kirchler and Maciejovsky (2002) above. To recall, the authors manipulated dividend information (complete and no information) and public signals about the market participants' average price prediction (precise, vague, no signal). First, the authors measured participants' risk aversion. In the main experiment, participants made bids and asks in an anonymous double auction to trade the risky assets. Before the market was opened, participants received either information about dividends distribution in the next round or no information and were asked to predict the average market price, state the 98% confidence interval, and state their confidence in the prediction on a 9-step scale. Consequently they received one of three public signals and then the market was opened. Participants were paid according to their trading performance in the experiment. The results suggested learning, because the observed average market prices came closer to expected prices with increasing experience (over time). The number of concluded contracts declined, the number of not accepted offers increased, and prices decreased over time. Surprisingly, Kirchler and Maciejovsky (2002) identified slight risk aversion, which was not significant among sessions. The results showed that information diffusion happens in the experiment. The participants in this experiment were highly overconfident in creating their confidence intervals (except for the first trading period) and their overconfidence increased over time. However, participants were generally not overconfident when overconfidence was measured on the 9-step scale (some were underconfident and some well-calibrated during some periods). Kirchler and Maciejovsky (2002) also found that trading volume was positively correlated with individual earnings and that subjects identified as overconfident in creating subjective confidence intervals earned significantly less than others. *The only question is whether the use of student participants is appropriate. They usually do not have experience with a market and especially with the creation of the confidence intervals in this environment (note that about one fifth of them were social science students). This fact could have contributed to the observed bias.*

Biais, Hilton, Mazurier, and Pouget (2005) experimentally measured the degree of overconfidence in judgment and self-monitoring. Student participants traded a single asset with uncertain dividend paid at the end of the trading game with asymmetric information and with permitted short-selling. Participants received heterogeneous private signals before trading. The asset was traded in oral double auction and call auction. The authors measured overconfidence outside the asset market. Participants were told that their final grade would reflect also the results in the experiment. Biais et al. (2005) found that their participants were greatly overconfident (only 64% of all confidence intervals contained the true value). They asked participants to state a 90% confidence interval for each of 10 questions. The authors then used the overconfidence measure – the proportion of answer falling outside the stated range – in the regression analysis as an explanatory variable for performance and trading behavior. Participants were told that their final grade would reflect also their results in the experiment. Biais et al. (2005) found that their participants were greatly overconfident (only 64% of all confidence intervals contained the true value). Moreover, the results suggested that miscalibration (overconfidence) reduced and self-monitoring (the disposition to attend to social cues) enhanced trading performance. *The experiment in Biais et al. (2005) has several shortcomings. First and most important, overconfidence was measured artificially outside the market. This is not a very fortunate choice because e.g., Klayman et al. (1999) showed that overconfidence differs over domains. Second, the incentives might not have been sufficient. Third, the authors did not control for risk aversion which might play a role in experiments involving uncertainty. Fourth, similar to the previously reviewed paper, the use of student participants may be problematic.*

These researchers identified overconfident behavior in experimental asset markets when confidence is measured by confidence intervals (e.g., Kirchler and Maciejovsky, 2002). Remarkably, the results suggest that overconfidence increases with experience. However, when measured directly on a nine-step scale (point estimates), Kirchler and Maciejovsky (2002) identified their participants as well-calibrated. In addition, the studies reviewed above showed convergence of prices to the optimal price.

As we have discussed above, overestimation is a common result in the literature when confidence intervals are used; it seems that either people really are badly calibrated when asked for confidence interval estimates, or this measure is not representative for them, or they have little experience with stating confidence intervals. However, auctions seem to be an effective means of obtaining equilibrium results nevermind the overestimation in confidence intervals. This is most probably due to the nature of the task – feedback in auctions.

3.4.6 Willingness to sell/buy

Willingness to buy and to sell is another paradigm used especially in experimental asset markets. This paradigm refers to the situation when price for which people are willing to buy a self-selected good (or asset) or willing to sell a self-selected good (or asset). Overconfidence (or overestimation) is in this case measured by the difference between the real (rational) and stated price. Thus, unlike the case of auctions, here people set only one limit price without further negotiation, and therefore the WTB/WTS paradigm lacks feedback to achieve convergence.

Fellner, Gueth, and Maciejovsky (2004) defined the illusion of expertise as “reluctance to give up, or eagerness to obtain, the individually selected portfolio in favor of an equally good alternative one” (p. 358). This can be regarded as overconfidence about one’s own expertise. The authors conducted an experiment with students who made financial decisions in 6 sessions with 12 participants each. First, participants completed 7 multiple choice decision tasks (analytical decision tasks and financial knowledge questions) and rated their experience and expertise. Second, in each of 2 identical periods they were randomly assigned to groups of four and each participant could invest an endowment into 4 risky assets. Future prices of assets were state-contingent with equal probabilities of three outcomes. All assets were perfectly positively correlated. Participants chose an expert from their group. A random price mechanism was used to elicit portfolio evaluations. Participants were also randomly assigned to willingness-to-pay (or willingness-to-accept) treatments and asked to choose maximum purchase (or minimum selling) price p from interval (-100,100) at which they were willing to switch from a self-selected to an alternative portfolio (expert or average group portfolio). Then, their p was compared to randomly chosen p*; if p>=p* that person had to buy his/her selected portfolio for p*; otherwise he received the alternative at no cost (or if p<=p* that person had to switch to the alternative portfolio and received p*; otherwise he kept the self-selected portfolio without compensation). According to the definition, illusion of expertise prevails if one is willing to pay positive prices to keep his portfolio (requires positive compensation for giving up). Future prices were then randomly determined and payoffs computed. Participants were paid proportionally to their performance in the experiment. In the decision task, only 58.54% of answers were correct and self-declared expertise was positively correlated with the number of correct answers. By choosing an expert, people mostly relied on self-ratings and answers to financial knowledge questions. The distribution of observed prices was skewed to the left, suggesting a systematic preference for the self-selected portfolios (even if this decision reduced the subject's payoff). More than half of the participants exhibited the illusion of expertise in both alternative portfolio treatments. In general, individuals were less willing to pay than willing to accept. In addition, 58% of individuals were classified as being constantly calibrated over the experiment. Comparing the results of the occurrence of positive prices for each individual, the authors concluded that the illusion of control occurs systematically and seems to be individually stable. *Fellner et al. (2004) used student participants and yet, given the difficulty of this task (portfolios and random price mechanism), it might have been the case that participants did not fully understand the experiment.*

Similarly, Dittrich, Gueth, and Maciejovsky (2005) investigated overconfidence in an investment setting. Student participants made investment decisions into one risky asset (in the first experiment) and into two risky assets (in the second experiment). They were paid according to their performance in the experiment. In the first experiment, participants made four successive investments of the endowment into four different risky assets. After each investment, participants determined limits for substituting their own portfolio with an alternative (optimal, higher than optimal, and lower than optimal – participants were not told this). Then, the random price mechanism described above (with price range (-49, 50)) was used to determine the payoffs. The results revealed that most investments were prominent shares of 100. Observed investments were on average 7.5 higher than optimal, and aggregate investments did not improve over time. The data indicated significant WTA/WTP disparity (the limit for WTA was higher than the limit for WTP). The median limit price for each of three alternatives was positive (for the optimal investment significantly lower), indicating overconfidence. Moreover, participants were more likely to be overconfident if the difference between the chosen and optimal investment was bigger. In the second experiment, participants had the possibility to invest into two risky assets (uncorrelated in one treatment and correlated in the other treatment). The results were similar to the results of the first experiment. *Dittrich et al. (2005) used student participants, who might not well understand the task, namely the random price mechanism.*

These results suggest disparity of willingness to pay (buy) and willingness to accept (sell) (Dittrich et al., 2005). People are generally willing to accept a higher price than willing to pay, even in situations where they shouldn't be willing to pay or accept anything (e.g., Fellner et al., 2004).

This bias might be connected to the underestimation of the endowment effect (as Van Boven, Loewenstein, and Dunning, 2003 suggested) as well as to the risk aversion that has not been investigated in this setting yet. However, Plott and Zeiler (2005) offered an alternative explanation, employing subject misconceptions stemming from the preference elicitation method. The authors showed that it is subject misconceptions resulting from the use of special mechanisms required to elicit valuations that causes the observed WTA/WTP gaps. Plott and Zeiler (2007) tested the endowment effect theory and other alternative explanations and showed that these cannot explain data from WTA/WTP experiments. In addition, List (2004) showed that consumers are able to learn to overcome the endowment effect.

3.4.7 Information

Information plays an important role in economics. No wonder experimental economists have incorporated various kinds of information into their research. Researchers investigate information processing and information acquisition in consumer research, finance, and forecasting. Since information builds the basis of every environment, this paradigm is one of the most diversified. In this section we review various information implementations in experimental studies and the corresponding results. We concentrate on studies that alter types or strength of information and investigate their impact on calibration.

Nelson, Bloomfield, Hales, and Libby (2001) were interested in the effect of information strength and weight on trading behavior in financial markets. In the first experiment, students traded artificial securities in laboratory financial markets. All participants in one market received a high-strength, low-weight 3-flip signal; all participants in the other market received a low-strength, high-weight 17-flip signal. Each coin (heads-biased and tails-biased) was flipped 3 or 17 times and placed into one of two buckets depending on the result. The value of each security was determined according to the proportion of heads-biased coins in the particular bucket. Each group of three investors traded 5 securities. To trade, each investor estimated the value of the security and chose linear demand (or supply). Then the computer found the market clearing price. In addition, investors were able to borrow money at no interest. Participants' payoff was based on their performance in the experiment. The results suggested overreaction to the information in the high-strength, low-weight 3-flip signal and underreaction to the information in the low-strength, high-weight 17-flip signal. The market was not able to correct for individual errors. In the second experiment, the authors combined investors with both types of signals into one market. In this case, 17-flip investors were underconfident relative to 3-flip investors (which were not significantly overconfident). In addition, prices were biased in the direction of overconfident 3-flip investors and a wealth transfer from 3-flip investors to 17-flip investors was observed. *Nelson et al. (2001) did not control for risk aversion which may play role in investing decisions under uncertainty. The subjects were also limited with the linear demand (supply) function. It is also not clear whether students were able to transform the signal mechanism into probabilities (frequencies) correctly.*

Subsequently, in the experiments of Nelson, Krische, and Bloomfield (2003), student participants had to decide how many shares to buy or sell. Investors first made their decisions on a case-by-case basis and then could modify their aggregate portfolio. A security cost $1 and had a value of $2 if the earnings growth was at least that predicted by an analyst and $0 otherwise. Participants also obtained recommended strategy, which was either to BUY (55% correct) or SELL (75% correct). One half of the recommended strategies in the experiment were BUY and one half were SELL. Participants were allowed to trade at least one and at most 50 securities of each of 30 firms. The participants were paid according to their performance in the experiment. The results revealed participants' confidence in the recommended strategy. The authors also identified an effect of accuracy of the signal (participants relied more on the recommended strategy when the signal was stronger) and of the aggregation (relied more on the recommended strategy when making portfolio-based judgments) on total trading profits. In the second experiment, the authors provided feedback about the profitability of the chosen strategies in the first 5 rounds. Then subjects were asked to estimate their performance on the subsequent 15 securities, in which they were provided with a BUY recommended strategy. The data from the no-feedback treatment showed a significant aggregation effect. In the feedback treatment, participants with negative feedback predicted their performance to be significantly worse than those with positive feedback. In addition, aggregation increased reliance on the trading strategy after negative feedback. *Nelson et al. (2003) did not control for risk aversion of their subjects, which might explain the observed behavior. One should also be careful in interpreting the difference in the signal strength because of the above mentioned WTB/WTS disparity (which happens with the equal signal strengths).*

Noeth and Weber (2003) experimentally investigated information aggregation. Six subjects were randomly ordered and in succession made predictions about the state based on the received signal. The signal was created in the following way: first, signal strength was randomly determined (50% weak, 50% strong); second, strong (weak) signal was correct with probability 4/5 (3/5). The prediction of every player was revealed to the others right after it had been made. Players were rewarded for a correct prediction. The result revealed that 107 out of 126 participants made better predictions than they would have made based only on private information. The data from the second predictor in the row exhibited overconfidence because people owerweighted their own weak signal and believed that the predecessor made more mistakes than he actually did. Also the third predictors' behavior exhibited overconfidence in the sense of overweighting their own information. However, the results also suggested that a certain degree of overconfidence was a better response to others' overconfident behavior. *Noeth and Weber (2003) used probabilities in their experiment. However, it has been shown that people work better with relative frequencies than with probabilities, especially in solving problems involving Bayesian updating.*

In a subsequent paper, Kraemer, Noeth, and Weber (2006) introduced costly information acquisition into the experimental design. Thus, the private information was not free anymore but individuals could buy it at a fixed price. Participants were paid according to their performance in the experiment. The results showed that participants bought 28% more information than Bayesian updating would suggest. Participants overestimated the value of the information especially at stage three, where they bought information in 41%

of cases after observing consistent public signals in the previous 2 stages (Bayesian updating would suggest that the value of information at this stage was much lower than the cost). In addition, the authors identified a depth of reasoning of level 2 (people were able to think what the predecessors thought but not further). *Kraemer et al. (2006), like Noeth and Weber (2003), did not provide information in frequencies but in probabilities.*

In sum, the results showed that in financial decisions people overreact to high-strength, low-weight information and underreact to low-strength, high-weight information (Nelson et al., 2001). As for information aggregation, people overestimate their own signals (e.g., Noeth and Weber, 2003). In other experiments, where information is costly and without aggregation, people overestimate the value of information, especially in latter stages where a higher level of Bayesian reasoning is needed (Kraemer et al., 2006). However, in some cases a certain degree of overconfidence is a better response to others' overconfident behavior (Noeth and Weber, 2003).

Since information was investigated in very different environments, it is difficult to draw clear conclusions for this paradigm. The main message might be that the quality as well as the quantity of information has a significant effect on calibration. Therefore, in studies interested in calibration (quality of judgments) researchers should use information that is the most representative for the situation under investigation. Obviously, one should investigate the impact of various kinds of information to find out how to improve calibration in various situations.

3.4.8 Assessment of others

We also identified a small body of literature investigating assessment of others and confidence in the accuracy of this assessment. Beliefs about others might be important, for example, in consumer research. Far more extensively investigated is the relative self-assessment that we reviewed in section 2.3. Yet even in relative self-assessment the assessment of others also plays a significant role because one has to assess others in order to estimate one's own percentile rank within the group. Below we briefly summarize the main findings of studies on calibration in assessment of others, where people assess actions, tastes, or calibration of others.

Shore, Adams, and Tashchian (1998) concentrated on the accountability issue of assessment of others. In the experiment, student subjects ("supervisors") were asked to evaluate the performance of other student subjects ("subordinates") on a clerical task – looking up catalogue numbers and prices of 40 items and computing a 15% discount for each price. The supervisors then marked the incorrect answers according to the master key. Then, they received additional information: knowledge of self-assessment of the subordinate (high or low – supposedly completed by the subordinate), evaluation purpose (for developmental feedback or for determining the likelihood of getting a job), and feedback target (face-to-face feedback to the subordinate or explain evaluation to a professor). After receiving the information, the supervisors evaluated the performance of their subordinates on a 6-point rating scale. Participants were not rewarded for their participation. The results showed that supervisors evaluated their subordinates significantly more positively when they received high self-assessment information. The authors also identified a significant interaction effect between the evaluation purpose and feedback target – when accountable to the organization (professor), there was less ratings inflation when ratings were used for administrative decisions (as hiring) than for developmental feedback and vice versa when feedback was meant to be for the subordinate. *Shore et al. (1998) neither motivated their subjects to perform as well as possible in the clerical task nor to provide as precise an evaluation as possible. Representativeness of stimuli used in the experiment is also questionable even though it is not directly connected with the evaluation method. Despite insufficient incentives and probably also representativeness, the authors showed that design and implementation corresponding as closely as possible to reality is very important in laboratory experiments and that deviation might lead to different results.*

Brenner, Koehler, Liberman, and Tversky (1996) investigated the estimates of actions of others and calibration in these estimates. Specifically, student participants had to rate themselves in terms of three bipolar dimensions (extrovert or introvert, analytical or intuitive, adaptive or decisive). Then participants were asked to complete 50 binary-choice questions involving dispositions, preferences, and behavior and to choose between 28 pairs of potential occupations. The authors then chose the two most common profiles (extrovert-intuitive-decisive and introvert-analytical-decisive) and another group of students made several predictions about these profiles. They received a flat participation reward. One group of participants was asked to estimate the predictions of a randomly selected individual from the group with the particular profile. In addition, they were asked to state their half-range confidence in each answer. Another group of participants first predicted the answer of the majority of target subjects and then stated the percentage estimate of those who selected that answer. A different group estimated the percentage of their peers that would choose a given alternative for each question regardless of the profile (base rate estimates). Brenner et al. (1996) also measured the representativness of both personality profiles - another group had to decide for each question, which of the two alternatives is more representative for a given profile. The results revealed that subjects did not make significantly different judgments of confidence in the individual and aggregate condition. Both groups exhibited overconfidence. The results also suggested that participants were successful in estimating the answers of a given profile. However, since the correlation between the targets' responses and base rate is higher than the correlation between targets' responses and predictions, participants did not use their knowledge of the base rate information appropriately. *Brenner et al. (1996) motivated their subjects only with flat incentives, which might not have been sufficient.*

Gershoff and Johar (2006) were interested in calibration of friends' knowledge. In the first experiment, the authors examined estimates of a friend's knowledge of one's own tastes (movies) and the friend's general-knowledge domain (capitals of US states). Participants first rated 40 movies (on a seven-point scale) and then estimated how accurate their friend would be (0-100 %) in estimating their movie tastes and how they would perform on a general-knowledge question quiz. Participants overestimated their friends' general knowledge as well as knowledge about their tastes in movies;

overconfidence in the latter was significantly higher for those who had an involved relationship. In the second and third experiments, the authors wanted to demonstrate that the motivation to maintain close relationships underpinned the enhanced overestimation of movie knowledge. Therefore, in both studies participants played both roles. In the third study, one half of the participants received feedback about their estimates and got a chance to correct them. Gershoff and Johar (2006) found support for their motivational hypothesis of enhanced overconfidence of personalized knowledge under high relationship involvement. In addition, the authors showed that participants believed that their close friends knew a lot about them and that they were more responsive to positive than negative feedback. *Gershoff and Johar (2006) used a flat incentives scheme in their experiment. Moreover, the choice of friends is a bit problematic.*

The results of this paradigm suggest overconfident behavior (e.g., Brenner et al., 1996). However, each of the reviewed papers investigates a different issue. People seem to be overconfident in their assessments of others and in others' assessments. This overconfidence is enhanced if there is a close relationship between people (Gershoff and Johar, 2006). This means that information does not seem to improve calibration because one can expect that close friends have better information about each other. Moreover, people do not use the available information sufficiently (Brenner et al., 1996). On the other hand, as in studies using other paradigms, feedback plays an important role here and lowers overconfidence (e.g., Shore et al., 1998).

The research in this paradigm mainly lacks incentives. In addition, attention should be paid to the representativeness of stimuli (questions, choices) as well as subjects used in the experiments. Assessment of others seems to depend on evaluation purpose and feedback target. Better understanding of this paradigm would help better understand the process of one's relative ranking (e.g., Kruger and Dunning, 1999) because if one wants to assess one's own position among group members correctly, she has to correctly assess her own performance as well as the performance of others. But note that in this case people are not assessing their "competitors" as is usual in studies involving the unskilled-and-unaware problem. Therefore, we should be careful when interpreting the outcomes and conclusions.

3.4.9 Self-awareness questions

Self-awareness questions are used to investigate the relative standings of individuals within a group. This paradigm essentially combines the assessment of others and absolute self-assessment, because in order to be able to place oneself within a group, it is necessary to know one's own performance and the performance of others (or at least their distribution). We have already reviewed research on this paradigm conducted in psychology. To recall, BTA and WTA effects were refined to the unskilled-and-unaware problem – the worse performing subjects (the unskilled) underestimate their relative standing, while the very skilled ones overestimate their relative standing, but less so. These patterns as well as other issues were also investigated in economics.

Ehrlinger, Johnson, Banner, Dunning, and Kruger (2008) replicated their earlier results (e.g., Kruger and Dunning, 1999), addressing various published critiques of their work (Ackerman, Beier, and Bowen, 2002; Burson et al., 2006; Krueger and Mueller, 2002; Krueger and Funder, 2004), and other concerns such as questions regarding the impact of financial incentives. The authors conducted 5 studies in various environments, with various tasks, and with various incentives. The tasks were the following: predictions of exam results; of the results of a regional debate tournament at Cornell University; of the results of a test of gun knowledge and safety on a Trap and Skeet competition; of the results on a multiple-choice test of logical reasoning ability, and another test of logical reasoning ability. The incentives in these tasks were as follows: extra course credit for participating; nothing; $5; extra course credit for participating; and extra course credit for participating, respectively. In all studies, subjects had to estimate their own score, average score, and their own ranking. Participants were mostly students (except for the participants of the Trap and Skeet tournament). The results essentially confirmed the results of previous studies – the existence of the unskilled-and-unaware problem. In addition, the results rejected alternative explanations of the unskilled-and-unaware problem, such as insufficient incentives or artificial situations. Accountability of subjects' estimates even worsened their calibration. The authors corrected for lack of reliability of measures in their analysis. *Ehrlinger et al. (2008) did not use representative stimuli in every task of their experiments. In addition, as we showed in Chapter 1 of this dissertation, the lack of insight might have been caused by insufficient information about the group participants, which was available for this kind of assessment.*

Niederle and Vesterlund (2007) were interested in the difference of self-selection of women and men into a competitive environment. Participants of their experiment had to, within five minutes, add up sets of five two-digit numbers without use of calculators. The experiment was computerized and they were informed about the correctness of each answer before adding up the next five numbers. Participants were divided into groups of four people and performed the task four times. Student participants were paid a participation fee as well as additional payment according to performance, but only for one randomly selected task: 50 cents for each correct answer, tournament – $2 for each correct task for the best performer, choice between 50 cents for each correct answer and tournament (comparing one's own performance on task 3 with performance of others on task 2 performance of others), and choice between 50 cents for each correct answer or tournament applying to task 1 results. The results did not reveal any gender difference in performance. However, there was a significant difference in selecting the tournament (competitions) – men entered the tournament twice as often as women. The authors concluded that this was not only because men are more overconfident but that they also have more preferences for tournaments. *Niederle and Vesterlund (2007) used representative stimuli in their experiments and also offered an alternative explanation of overconfident behavior of men – preference for tournament.*

Hoelzl and Rustichini (2005) defined overconfidence in a general way – when a majority of people estimate their skills or abilities to be better than the median. The authors used hard and easy tasks with and without monetary incentives in their experiment, for a 2x2 design. The task was to complete a gap-filling exercise of 20 sentences in a test of knowledge of vocabulary (LEWITE test). In each task participants had to choose two words from 7-9 alternatives. However, the authors did not measure confidence directly they let a group of people vote by majority rule for one of two payoff conditions (performance or lottery). In the performance condition only those whose performance was in the upper half of the results of all participants would be paid. In the lottery condition everyone had a 50% chance of winning money. In addition, subjects were asked for estimates of their own and group average performance. Participants (mostly students) in the monetary condition could have won money. The average vote was 55% in favor of the test. From the data it followed that increasing task difficulty as well as not offering monetary incentives encouraged voting for lottery. The authors found that the behavior significantly changed with task difficulty only if money was offered (overconfidence if an easy, familiar task; underconfidence if a non-familiar task). *The basic problem of Hoelzl and Rustichini (2005) is their strategy for recruiting students, which might have caused a sample selection problem. Subjects for this experiment were recruited in classrooms and the local cafeteria until a sufficiently big group was gathered. Note that this is the case where relative self-assessment matters and therefore also the subject pool plays a role. This fact magnifies the improper recruitment strategy.*

All studies that we reviewed under the market-entry games paradigm in Section 4.4 should be included in this section because when making market-entry decisions, one has to assess oneself as well as others and consider if one would perform well enough to reap a positive profit. We will not repeat the results at this point, but highlight that overconfidence (excessive entry) was the most frequent result, though only for easier tasks, while for harder tasks underconfidence was observed.

In this paradigm, various measures were used to measure people's relative self-assessment – direct percentile ranking, lotteries, tournaments, and market-entry games. The first measurement is the most refined and confirms the finding from psychology – that the unskilled-and-unaware problem seems to be persistent (Ehrlinger et al., 2008). Measuring self-assessment through a choice between lottery and performance-based payment reveals overconfidence for easier tasks and underconfidence for harder tasks (Hoelzl and Rustichini, 2005). The difficulty argument also applies to market-entry games (e.g., Moore et al., 2007). Finally, the evidence suggests that men are more competitive (and overconfident) than women (Niederle and Vesterlund, 2007).

In Chapter 1 of this dissertation we pointed out the importance of the subject pool used in these kinds of experiments and the positive effect of information on calibration (experimentally supported in the Chapter 2 of this dissertation). Therefore, we conclude that the question of the subject pool should be taken into account, especially in experiments involving comparison of members of the group.

3.5 Conclusion and discussion

In this review, we have focused on overconfidence – one of the most frequently and widely investigated biases in recent decades. The aim of this survey was to categorize, review, and evaluate experimental studies on overconfidence and self-assessment in business, economics, and finance, to enumerate the shortcomings of the studies in each of the paradigms that we identified, and to prioritize promising future research ideas.

We started by reviewing the basic findings and issues from psychology where general-knowledge questions, sensory discrimination tasks, and personal assessment questions are the three dominant paradigms.

Overconfidence seems to prevail in psychology experiments that use general-knowledge questions. This finding was, with the introduction of task difficulty, refined by the discovery of the so-called hard-easy effect (overconfidence appearing in hard-item samples and underconfidence in easy-item samples; e.g., Griffin and Tversky, 1992). Several methodological problems that might artificially contribute to the overconfidence bias were subsequently pointed out (regression-to-the-mean – e.g., Pfeifer, 1994; error in the judgment process – Erev et al., 1994; representative design – Gigerenzer et al., 1991). Several ways to improve calibration were also identified (feedback – e.g., Arkes et al., 1987; discussion of choices/results – Pulford and Colman, 1997; counter-reasoning – Koriat et al., 1980). Importantly, Juslin et al. (2000) showed that overconfidence (as well as the hard-easy effect) in general-knowledge questions seems to emerge typically in studies that do not use representative stimuli (i.e., samples in which the stimuli are not sampled in a random manner).

In contrast, experimental studies employing sensory discrimination tasks appear to elicit exclusively underconfident behavior that seems immune even to feedback (e.g., Bjoerkman et al., 1993). The substantial difference in calibration in general-knowledge questions and in sensory discrimination tasks seems to be caused by the different types of error involved in the confidence judgment processes used in these two paradigms (Juslin and Olsson, 1997).

Furthermore, studies employing personal assessment questions identified the so-called Better-Than-Average (BTA) as well as Worse-Than-Average (WTA) effects in relative self-assessment (e.g., Svenson, 1981). These findings were later extended to the unskilled-and-unaware problem (e.g., Kruger and Dunning, 1999) which refers to a difference in the calibration of ranking of the skilled (underestimation) and the unskilled (overestimation). As with general-knowledge questions, overconfident behavior was observed more frequently than underconfident behavior. Personal assessment questions tie the question of overconfidence to problems of calibration, Bayesian updating, incentives, availability of information (e.g., Moore and Healy, 2008), and feedback.

Thus, the main issues so far identified in experimental studies in psychology are representativeness of stimuli (e.g., Juslin et al., 2000), representation issues (e.g., Brenner et al., 1996), and presence of feedback (e.g., Petrusic and Baranski, 1997).

In addition, we reviewed the most important issues in experimental studies in business, economics, and finance. In our view, these are the use of financial (or other) incentives (e.g., Camerer and Hogarth, 1999; Rydval and Ortmann, 2004), external/ecological validity of experiments (e.g., Harrison and List, 2004), the subject pool used (see Chapter 1 of this dissertation), and possible alternative explanations for the results (e.g., risk aversion).

We used the *Econlit* and *Web of Science* databases to create, using a clear and replicable selection rule, a non-opportunistic set of experimental studies from business, economics, and finance involving overconfidence or self-assessment issues. In these studies, we identified nine distinct paradigms: General-knowledge questions, Confidence intervals, Forecasting, Market-entry games, Auctions, Willingness to sell/buy, Information, Assessment of others, and Self-awareness questions. It seems fair to say that the paradigms addressing issues of overconfidence are more heterogeneous in business, economics, and finance than in psychology. We organized our review of the experimental studies according to these paradigms, paying attention to the seven issues identified above. We also made suggestions for further research wherever applicable.

A propensity towards overconfident behavior was identified in the literature investigating calibration with *general-knowledge questions* (e.g., Mahajan, 1992). Calibration in *general-knowledge questions* seems to improve with feedback (e.g., Arkes et al., 1987) and counterfactual reasoning (Mahajan, 1992). It has also been shown that frequency responses return more accurate confidence estimates than probability responses (Price, 1998). Moreover, the evidence seems to suggest that experiments with financial incentives generate different results than those without (Hoelzl and Rustichini, 2005). Wallsten et al. (1993) showed that verbal and numerical statements of confidence differ, however neither of them is better. Alternative ways of measuring people’s confidence in their performance were also used: Hoelzl and Rustichini (2005) used the choice between lottery- and performance-based rewards. Notwithstanding the important results from psychology, researchers in economics continue to underestimate the importance of representative sampling of general-knowledge questions. We conclude that researchers should be aware of the (un)representative nature of their stimuli (including representative subject samples) used in their experiments and the amount of information.

Results from studies on *confidence intervals* suggest much stronger overconfidence (e.g., Oenkal et al., 2003) than those with conventional measures of confidence such as directly stating confidence in point estimates (subjective probability that the estimate is correct). It was also shown (e.g., Du and Budescu, 2007) that overconfidence is present mainly in high-confidence confidence intervals (e.g., 90%) while underconfidence prevails in low-confidence confidence intervals (e.g., 50%) and no miscalibration in mid-confidence confidence intervals. Yaniv and Foster (1995, 1997) suggested that people make a trade-off between accuracy and informativeness when stating interval forecasts. Moreover, experts seem to be able to create more accurate confidence intervals than non-experts (Oenkal et al., 2003) although non-experts clearly learn and they do so relatively quickly (e.g., Cesarini et al., 2006). The effect of feedback (Bolger and Oenkal-Atay, 2004) and even pure repetition of confidence interval estimation (Cesarini et al., 2006) on calibration was found to be significantly positive. Klayman et al. (1999) showed that overconfidence is not linearly related to task difficulty, yet it differs across domains. We argue that it is inappropriate to measure overconfidence outside the task under consideration (e.g., Biais et al., 2005). We also suggest, as others have done before us (e.g, Cesarini et al., 2006), that people might not be familiar with the notion of confidence intervals, as few people use confidence intervals in their daily lives where they more often express their confidence in the form of statements or choices. The positive impact of simple repetition on calibration (e.g., Cesarini et al., 2006) would seem to support this claim.

Experimental research in *forecasting* suggests, up to this point, mostly overconfident behavior, especially if confidence intervals are predicted (e.g., Bolger and Oenkal-Atay, 2004). However, confidence in the correctness of confidence intervals, or confidence in point estimates, is not necessarily excessive (e.g., Oenkal et al., 2003). Overconfidence is usually decreased through feedback or practice (e.g., Kelemen et al., 2007). Economic experts exhibit higher confidence and surprisingly lower performance in forecasting tasks than do students, and consequently economic experts seem to be more overconfident (e.g., Thomson et al., 2003). In contrast, weather forecasters exhibit much lower overconfidence than economists (e.g., Tyszka and Zielonka, 2002). We argue that the problem in forecasting studies might be the method used to measure overconfidence. Specifically, we argue that predictions should be compared with the best possible prediction given the available information and not with the actual result (which might actually be biased in one direction). We also suggest that unrepresentativeness of the stimuli materials, which are usually selected (non-randomly) by experimenters, might be the reason that experts perform worse than students. Alternatively, the lack of (sufficient) incentives for experts might also be a reason for this surprising result.

In *market-entry games* it was shown that if the games are played repeatedly, people tend to be well-calibrated in their decisions and convergence is, with more information, achieved faster (e.g., Erev and Rapoport, 1998). After introducing skill-dependent payoffs (Camerer and Lovallo, 1999), people were identified as overconfident in their own skills in easy tasks (markets), resulting in excessive entry, and underconfident in difficult tasks (markets), resulting in insufficient entry (e.g., Moore et al., 2007). In the case of skill-dependent payoffs in market-entry games, feedback does not always overcome miscalibration (e.g., Moore and Cain, 2007). An important caveat regarding the research on market-entry games is that researchers do not pay attention to insights from psychology regarding, for example, representativeness of stimuli (or at least they do not report how the questions were selected; e.g., Hogarth and Grieco, 2004). We also note that the dependence of over- and underconfidence on task difficulty is at odds with Juslin et al. (2000), who showed that task difficulty does not influence overconfidence in absolute self-assessment on two-alternative general-knowledge questions. The results of Elston, Harrison, and Rustroem (2005) suggest that task difficulty and risk attitudes have no effect on the probability of entry but that the predicted number of entrants and the subjects' confidence in their skills do. These results suggest two possible explanations of reference group neglect (identified in Camerer and Lovallo, 1999): either it might be a laboratory (experimental) artifact, or it might apply only to a selected group of people (e.g., the so-called part-time entrepreneurs in Elston et al., 2005), or both. Moreover, we claim that risk aversion is likely to play an important role in market-entry games as these represent a risky situation involving uncertainty (about the absolute/relative abilities of others) and risk aversion measures should therefore be included into this class of experiments.

Auctions were typically used in experimental asset markets. In experimental asset markets, overconfident behavior seems to prevail when confidence is measured by confidence intervals (Kirchler and Maciejovsky, 2002). However, when measured directly on a nine-step scale (confidence in the correctness of the stated price), people seem to be well-calibrated (Kirchler and Maciejovsky, 2002). Remarkably, overconfidence (measured by confidence intervals) seems to increase with experience. Even though overconfidence prevails if auctions are used, auctions seem to be an effective instrument to generate equilibrium results (convergence to the equilibrium price). This convergence is most probably due to the provided information (feedback) that is present in auctions (in the form of asks and bids of other participants). We therefore conjecture that low calibration in confidence intervals used in auctions might be caused by people's lack of familiarity with this measure.

In studies investigating the *willingness to pay (buy)* and *willingness to accept (sell)* we identified a disparity of these two measures (e.g., Dittrich et al., 2005). People are generally willing to accept a higher price than they are willing to pay, even in situations where they shouldn't be willing to pay or accept anything (e.g., Fellner et al., 2004). This "disparity" bias might be connected to the underestimation of the endowment effect (as Van Boven, Loewenstein, and Dunning, 2003 suggested) as well as to the risk aversion that has not been investigated in this setting yet. Plott and Zeiler (2007, see also 2005), however, tested the endowment effect theory and other alternative explanations. Their results suggest that neither endowment effect theory nor other alternative explanations can explain data from WTA/WTP experiments in a satisfactory manner (see also Isoni, Loomes, and Sugden, 2008). In addition, List (2004) showed that consumers are able to learn to overcome the endowment effect.

The role of *information* was experimentally investigated with results showing that in financial decisions people overreact to high-strength, low-weight information and underreact to low-strength, high-weight information (Nelson et al., 2001). As regards information aggregation, the findings by Noeth and Weber (2003) suggest that people overestimate their own signals. The results by Kraemer et al. (2006) indicate that in experiments with costly information and without aggregation, people overestimate the value of information, especially in latter stages where a higher level of Bayesian reasoning is needed. On the other hand, small overconfidence might be a better response to others' overconfident behavior in some cases (Noeth and Weber, 2003). Once again, we observe that in this paradigm insufficient attention was paid to representativeness of the stimuli used in the experiments (e.g., Shore et al., 1998). We note that such information should be used in studies investigating calibration (quality of judgments) that is the most representative for the situation under investigation. Obviously, one might also investigate the impact of various kinds of information to find out how to improve calibration in various situations.

The results speaking to the *assessment of others* paradigm suggest that people are overconfident in their assessments of others and in others' assessments (e.g., Brenner et al., 1996), the more so the closer the relationship between people is (Gershoff and Johar, 2006). Feedback seems to lower overconfidence, although people do not use the available information sufficiently (e.g., Shore et al., 1998). We point out that studies investigating this paradigm often use insufficient financial incentives (e.g., Shore et al., 1998). In addition, more attention should be paid to the representativeness of stimuli (questions, choices; e.g., it is not clear in Shore et al., 1998). We also note that better understanding of this paradigm would help better understand the process of one's relative ranking (e.g., Kruger and Dunning, 1999) because if one wants to assess one's own position among group members (rank) correctly, she has to correctly assess her own performance as well as the performance of others.

Various measures were used to measure people's *relative self-assessment.* First, direct percentile rankings seem to result in the unskilled-and-unaware problem (e.g., Ehrlinger et al., 2008). Second, the choice between lottery and performance-based payments suggest overconfidence for easier tasks and underconfidence for harder tasks (Hoelzl and Rustichini, 2005). Third, overconfidence (excessive market entry) for easier tasks and underconfidence (insufficient market entry) for harder tasks were found in market-entry games (e.g., Moore et al., 2007). Fourth, research on tournaments suggested that men are more competitive (and more overconfident) than women (Niederle and Vesterlund, 2007). In Chapter 1 of this dissertation we pointed out the importance of the subject pool used in these kinds of experiments and the positive effect of information on calibration (see chapter 2 of the present dissertation for experimental evidence of this claim). We therefore conclude that the question of subject pool should be taken into account, especially in experiments involving comparison of the members of the group.

As we have shown above, the term *overconfidence* denotes a bias that manifests itself in various forms. The basic three forms are overestimation of one's actual performance, overplacement of one's performance relative to others, and excessive precision in one's beliefs (as classified in Moore and Healy, 2008). Overconfidence varies not only among these forms but also among different task domains if the same form of overconfidence is measured. Note that, in many situations, people were identified as well-calibrated. Therefore, when referring to overconfident behavior, one should specify both the domain and the elicitation details.

In our review, we concentrated on the existence of over- and underconfidence and how these miscalibrations can be reduced. In some cases, however, slight overconfidence might be desirable – either as strategic behavior or as socially desirable behavior. For example, Noeth and Weber (2003) showed that slight overconfidence is a better response to the behavior of others than is "optimal" behavior. Overconfidence can be beneficial also in other situations. In a situation in which all agents are risk averse, it could happen that no one is willing to enter a market. Overconfident behavior, by pushing some agents over the threshold to market-entry, could remedy this problem.

In our literature review, we pointed out the main shortcomings of existing experiments or experimental classes. Despite the clear message from psychology (where representativeness of stimuli was shown to be the key factor that caused the seemingly persisting overconfidence bias to disappear, see e.g., Juslin et al., 2000), representativeness of stimuli is probably the most ignored issue in experimental studies from business, economics, and finance. Representativeness of stimuli means random sampling of questions and of alternative answers (e.g., for general-knowledge questions) or other stimulus material used in these experiments. Use of non-representative stimuli might lead to misleading results by identifying biased behavior even if people are, in fact, well-calibrated (or, much better calibrated). In addition, the lack of financial incentives for participants might, due to insufficient motivation, artificially contribute to the observed biases.

We reviewed many studies that utilize general-knowledge questions as stimuli. We also reviewed studies that investigate confidence in skill-oriented tasks (e.g., summing numbers in Niederle and Vesterlund, 2007). However, there is no evidence of a direct comparison of calibration in confidence of the two big classes of tasks: general-knowledge questions and skill tasks (except for a comparison of sensory discrimination tasks and general-knowledge questions in, e.g., Juslin et al., 2003). Therefore, the relationship between calibration in general-knowledge oriented tasks and calibration in skill-oriented tasks remains an open question: What are the patterns of miscalibration in these two types of tasks and why do these patterns differ (if indeed they do)?

Feedback helps to improve calibration in many cases (or even simple repetition might be helpful in some cases, see e.g., Cesarini et al., 2006). Such improvement in calibration might be caused by natural learning and/or receiving additional information that was not available before. To distinguish between these two drivers of improvement in calibration, one could, in one treatment, provide all possible information at the beginning and, in another treatment, provide low information but with feedback. In situations where one expects people to be unfamiliar with the task/environment/measuring methods (as it was, for example, in the case of confidence intervals in Cesarini et al., 2006), feedback (or an alternative learning method) should be provided to reduce noise. Nevertheless, adequate information should be provided in all cases, except when the impact of various kinds of information is being investigated.

As already mentioned, the assessment-of-others paradigm is very important for relative self-assessment where one has to assess both oneself and others. Therefore, in order to fully understand the observed miscalibration in relative self-assessment, assessment of others should first be properly investigated. Specifically, first we should know what kind of information (and in what way) people use when assessing the performance of others. Second, one needs to know how the estimation of one’s own performance affects the estimates of others’ performances. And finally, one needs to know whether people use these two estimates in their own ranking estimation.

Market-entry games connect self-assessment with assessment of others, as well as with competitiveness of market participants (subjects) and therefore provide a very promising framework. The link between the unskilled-and-unaware problem and excessive/insufficient market entry is intriguing. Recall that Camerer and Lovallo (1999) identified excessive entry while Elston et al. (2005) found their subjects (2 out of 3 groups) well-calibrated. Note that Elston et al. (2005) used 5 subjects and market capacity equal to 1 while Camerer and Lovallo (1999) used 8 subjects and market capacities equal to 2, 4, 6, and 8. In the presence of the unskilled-and-unaware problem, we can expect insufficient entry for very low market capacities (only the very skilled should enter and these are, according to the unskilled-and-unaware problem, underestimating their rank). Similarly, for higher market capacities we can expect excessive entry (because the unskilled are overestimating their rank). Note that the switch between under- and overestimation is usually somewhere between the top 20-40th percentile. The unskilled-and-unaware problem could provide an additional explanation for the difference in results found in Camerer and Lovallo (1999) and Elston et al. (2005), besides the subject pool explanation discussed in Elston et al. (2005). Moreover, one could investigate what kind of feedback is the most helpful for achieving market rates closer to the equilibrium rates.

Alternatively, one could change the setup of market entry games so that the market capacity would not be fixed. The idea is that, instead of a fixed market capacity c, there would be a given profit P to be captured by market entrants whose payoffs would then depend not only on their relative ranking but also on their absolute performance (and absolute performance of all other entrants). Given the performance of the entrants $y_1, \ldots, y_t$, the threshold c could be determined so that $\Sigma(y_i - c) = P$. This setup has several advantages: First, we are not fixing the number of entrants, which makes the game more like real world situations. Second, not only relative ranking matters but also absolute performance, which also makes the game setup more realistic. Third, the results are less affected by errors in judgements (if two people with approximately the same performance enter, their profit/loss will be approximately the same, which is not so in the classic setup, where if one of them happens to be above the threshold and the other one below, their payoffs markedly differ). Fourth, it would be possible to quantify overconfidence, which was not the case in the classic setup where we could only observe the number of entrants and compare it to the market capacity. With this new setup, we might expect lower overconfidence rates, especially in cases when the performance of subjects is more even.

Our literature review shows how broad the range of situations is in which overconfidence seems to operate. Our review also suggests that many of these situations may be the result of experimental design and implementation idiosyncrasies, or be experimental artifacts; in short, they may not transfer from the lab to the field. Clearly, a barrage of unanswered questions concerning the alleged overconfidence bias remains to be answered.

References

Ackerman, P.L., Beier, M.E., and Bowen, K.R., 2002. What We Really Know about Our Abilities and Our Knowledge. *Personality and Individual Differences, 33 (4)*, 587-605.

Alba, J.W., Hutchinson, W.J., 2000. Knowledge Calibration: What Consumers Know and What They Think They Know. *Journal of Consumer Research, 27,* 123-156.

Andersson, P., Edman, J., and Ekman, M., 2005. Predicting the World Cup 2002 in Soccer: Performance and Confidence of Experts and Non-experts. *International Journal of Forecasting, 21,* 565– 576.

Angner, E., 2006. Economists as experts: Overconfidence in Theory and Practice. *Journal of Economic Methodology, 13 (1),* 1–24.

Arkes, H.R., Christensen, C., Lai, C., and Blumer, C., 1987. Two Methods of Reducing Overconfidence. *Organizational Behavior and Human Decision Processes, 39 (1),* 133-144.

Armantier, O., 2003. Estimates of Own Lethal Risks and Anchoring Effects. *Journal of Risk and Uncertainty, 32*, 37–56.

Baranski, J.V., Petrusic, W.M., 1994. The Calibration and Resolution of Confidence in Perceptual Judgments. *Perception & Psychophysics, 55 (4),* 412-428.

Baranski, J.V., Petrusic, W.M., 1999. Realism of Confidence in Sensory Discrimination. *Perception & Psychophysics, 61 (7),* 1369-1383.

Barber, B.M., Odean, T., 2000. Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors. *Journal of Finance, 55 (2),* 773-806.

Barber, B.M., Odean, T., 2001. Boys Will be Boys: Gender, Overconfidence, and Common Stock Investment. *The Quarterly Journal of Economics, 116 (1),* 261-292.

Biais, B., Hilton, D., Mazurier, K., and Pouget, S., 2005. Judgmental Overconfidence, Self-monitoring, and Trading Performance in an Experimental Financial. *Review of Economic Studies, 72,* 287-312.

Bjoerkman, M., Juslin, P., and Winman, A., 1993. Realism of Confidence in Sensory Discrimination: The Underconfidence Phenomenon. *Perception & Psychophysics, 54 (1),* 75-81.

Bolger, F., Oenkal-Atay, D., 2004. The Effects of Feedback on Judgmental Interval Predictions. *International Journal of Forecasting, 20 (1),* 29-39.

Bolger, F., Harvey, N., 1995. Judging the Probability that the Next Point in an Observed Time-series Will Be Below, or Above, a Given Value. *Journal of Forecasting, 14,* 597-607.

Brenner, L.A., Koehler, D.J., Liberman, V., and Tversky, A., 1996. Overconfidence in Probability and Frequency Judgments: A Critical Examination. *Organizational Behavior and Human Decision Processes, 65 (3),* 212-219.

Bruce, A.C., Johnson, J.E.V., 2001. Market Ecology and Decision Behaviour in State-Contingent Claims Markets. *Journal of Economic Behavior and Organization, 56,* 199–217.

Budescu, D.V., Erev, I., and Wallsten, T.S., 1997. On the Importance of Random Error in the Study of Probability Judgment. Part I: New Theoretical Developments. *Journal of Behavioral Decision Making, 10,* 157-171.

Budescu, D.V., Wallsten, T.S., and Au, W.T., 1997. On the Importance of Random Error in the Study of Probability Judgment. Part II: Applying the Stochastic Judgment Model to Detect Systematic Trends. *Journal of Behavioral Decision Making, 10 (3),* 173 – 178.

Budescu, D.V., Du, N., 2007. The Coherence and Consistency of Investors' Probability Judgments. *Management Science, 53 (11),* 1731-1744.

Burson, K.A., Larrick, R.P., and 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 (1),* 60–77.

Camerer, C., 1995. Individual Decision Making, Handbook of Experimental Economics. Princeton University Press.

Camerer, C., 2003. Behavioral Game Theory: Experiments in Strategic Interaction. Princeton University Press.

Camerer, C., Ho, T., Chong, K., 2003. Models of Thinking, Learning, and Teaching in Games. *American Economic Review, 93 (2),* 192-195.

Camerer, C., Hogarth, R.M., 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 and Organization, 61 (3),* 453-470.

Chuang, W., Lee, B., 2006. An Empirical Evaluation of the Overconfidence Hypothesis. *Journal of Banking and Finance, 30 (9),* 2489-2515.

Cochrane, J.H., 2001. Asset Pricing. Princeton University Press.

Compte, O., Postlewaite, A., 2004. Confidence-Enhanced Performance. *American Economic Review, 94 (5),* 1536-1557.

Daniel, K.D., Hirshleifer, D., and Subrahmanyam, A., 2001. Overconfidence, Arbitrage, and Equilibrium Asset Pricing. *Journal of Finance 56 (3),* 921–965.

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

Dittrich, D., Gueth, W., and Maciejowsky, B., 2005. Overconfidence in Investment Decisions: An Experimental Approach. *European Journal of Finance, 11 (6),* 471-491.

Du, N., Budescu, D.V., 2007. Does Past Volatility Affect Investors' Price Forecasts and Confidence Judgements? *International Journal of Forecasting, 23,* 497–511.

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

Dunning, D., Johnson, K., Ehrlinger, J., and Kruger, J, 2003. Why People Fail to Recognize Their Own Incompetence. *Current Directions in Psychological Science 12 (3),* 83–87.

Edwards, R.K., Kellner, K.R., Sistrom, C.L., and Magyaria, E.J., 2003. Medical Student Self-assessment of Performance on an Obstetrics and Gynecology Clerkship. *American Journal of Obstetrics and Gynecology, 188 (4),* 1078-1082.

Eggertsson, T., 1990. Economic Behavior and Institutions. Cambridge Surveys of Economic Literature.

Ehrlinger, J, Johnson, K., Banner, M., Dunning, D., and Kruger, J., 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., and Rutstroem, E.E., 2005. Characterizing the Entrepreneur Using Field Experiments. *Working Paper 05-30, Department of Economics, College of Business Administration, University of Central Florida*.

Erev, I., Rapoport, A., 1998. Coordination, ‘‘Magic,’’ and Reinforcement Learning in a Market Entry Game. *Games and Economic Behavior, 23,* 146-175.

Erev, I., Wallsten, T.S., and Budescu, D.V., 1994. Simultaneous Over- and Underconfidence: The Role of Error in Judgment Processes. *Psychological Review, 101 (3),* 519-27.

Fama, E.F., 1998. Market Efficiency, Long-term Returns, and Behavioral Finance. *Journal of Financial Economics, 49,* 283-306.

Fellner, G., Gueth, W., and Maciejovsky, B., 2004. Illusion of Expertise in Portfolio Decisions: An Experimental Approach. *Journal of Economic Behavior and Organisation, 55,* 355-176.

Frankenberger, K.D., Albaum, G.S., 1997. Using Behavioral Decision Theory to Assess Advertisement Recognition Tasks by Level of Difficulty. *Psychology and Marketing, 14 (2),* 145-162.

Friedman, D., 1998. Monty Hall's Three Doors: Construction and Deconstruction of a Choice Anomaly. *American Economic Review, 88 (4),* 933-946.

Gershoff, A.D., Johar, V.G., 2006. Do You Know Me? Consumer Calibration of Friends' Knowledge. *Journal of Consumer Research, 32,* 496-503.

Gigerenzer, G., Hoffrage, U., 1995. How to Improve Bayesian Reasoning Without Instruction: Frequency Formats. *Psychological Review, l02 (34),* 684-704.

Gigerenzer, G., Hoffrage, U., 1999. Overcoming Difficulties in Bayesian Reasoning: A Reply to Lewis and Keren (1999) and Meilers and McGraw (1999). *Psychological Review, 106 (2),* 425–430.

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

Griffin, D., Tversky, A., 1992. The Weighing of Evidence and the Determinants of Confidence. *Cognitive Psychology, 24,* 411-435.

Harrison, G.W., List, J.A., 2004. Field Experiments. *Journal of Economic Literature, 42,* 1009–1055.

Haun, D.E., Zeringue, A., Leach, A., and Foley, A, 2000. Assessing the Competence of Specimen-Processing Personnel. *Laboratory Medicine, 31,* 633–637.

Hertwig, R., Ortmann, A., 2001. Experimental practices in Economics: A Methodological Challenge For Psychologists? *Behavioral and Brain Sciences, 24,* 383-451.

Hirshleifer D., 2001. Investor Psychology and Asset Pricing. *Journal of Finance, 56 (4),* 1533-1597.

Hoffrage, U., Gigerenzer, G., Krauss, S., and Martignon, L., 2002. Representation Facilitates Reasoning: What Natural Frequencies Are and What They Are Not. *Cognition 84,* 343–352.

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

Hogarth, R.M., Grieco, D., 2004. Excess Entry, Ambiguity Seeking, and Competence: An Experimental Investigation. *Economics Working Paper 778, Universitat Pompeu Fabra*.

Isoni, A., Loomes, G., and Sugden, R., 2008. The Willingness to Pay-Willingness to Accept Gap, the "Endowment Effect", Subject Misconceptions, and Experimental Procedures for Eliciting Valuations: Comment. *Manuscript*, http://www.uea.ac.uk/eco/ecopeople/LoomesG.html

Juslin, P., 1994. The Overconfidence Phenomenon as a Consequence of Informal Experimenter-Guided Selection of Almanac Items. *Organizational Behavior and Human Decision Processes, 57,* 226-246.

Juslin, P., Olsson, H., 1997. Thurstonian and Brunswikian Origins of Uncertainty in Judgment: A Sampling Model of Confidence in Sensory Discrimination. *Psychological Review, 104 (2),* 344-366.

Juslin, P., Wennerholm, P., and Olsson, H., 1999. Format Dependence in Subjective Probability Calibration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25 (4)*, 1038-1052.

Juslin, P., Winman, A., and Hanson, P., 2007. The Naïve Intuitive Statistician: A Naïve Sampling Model of Intuitive Confidence Intervals. *Psychological Review, 114 (3),* 678–703.

Juslin, P., Winman, A., and Olsson, H., 2000. Naïve Empiricism and Dogmatism in Confidence Research: A Critical Examination of the Hard-Easy Effect. *Psychological Review, 107 (2),* 384-396.

Juslin, P., Winman, A., Olsson, H., 2003. Calibration, Additivity, and Source Independence of Probability Judgments in General Knowledge and Sensory Discrimination Tasks. *Organizational Behavior and Human Decision Processes, 92,* 34-51.

Kahneman, D., 1988. Experimental Economics: A Psychological Perspective. In R. Tietz, Wulf Albers, and Reinhard Selten, eds., Bounded Rational Behavior in Experimental Games and Markets. New York: Springer-Verlag, 11-18.

Kelemen, W.L., Winningham, R.G., and Weaver, C.A., 2007. Repeated Testing Sessions and Scholastic Aptitude in College Students' Metacognitive Accuracy. *European Journal Of Cognitive Psychology, 19 (4/5),* 689-717.

Keren, G., 1999. On The Calibration of Probability Judgments: Some Critical Comments and Alternative Perspectives. *Journal of Behavioral Decision Making, 10 (3),* 269 – 278.

Kirchler, E., Maciejovsky, B., 2002. Simultaneous Over- and Underconfidence: Evidence from Exprerimental Asset Markets. *Journal of Risk and Uncertainty, 25 (1),* 65-85.

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

Koellinger, P., Minniti, M., and Schade, C., 2007. "I think I can, I think I can": Overconfidence and Entrepreneurial Behavior. *Journal of Economic Psychology, 28 (4)*, 502-527.

Kraemer, C., Noeth, M., and Weber, M., 2006. Information Aggregation with Costly Information and Random Ordering: Experimental Evidence. *Journal of Economic Behavior and Organization, 59,* 423–432.

Krueger, J.I., Funder, D.C., 2004. Towards a Balanced Social Psychology: Causes, Consequences, and Cures for The Problem-Seeking Approach To Social Behavior and Cognition. *Behavioral and Brain Sciences, 27*, 313-327.

Krueger, J.I., Mueller, R.A., 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 (2),* 180-188.

Kruger, J., 1999. Lake Wobegon Be Gone! The "Below-Average Effect" and the Egocentric Nature of Comparative Ability Judgments. *Journal of Personality and Social Psychology, 77 (2),* 221-32.

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 (6)*, 1121-1134.

Kruger, J., Dunning, D., 2002. Unskilled and unaware--but why? A reply to Krueger and Mueller (2002). *Journal of Personality and Social Psychology, 82 (2),* 189-192.

Kogan, S., 2006. Distinguishing Overconfidence from Rational Best-Response in Markets. Available at SSRN: http://ssrn.com/abstract=891382.

Koriat, A., Lichtenstein, S., and Fischhoff, B., 1980. Reasons for Confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 107-118.

Kovalchik, S., Camerer, C.F., Grether, D.M., Plott, C.R., and Allman, J.M., 2005. Aging and Decision Making: A Comparison Between Neurologically Healthy Elderly and Young Individuals. *Journal of Economic Behavior and Organization, 58*, 79–94.

Lichtenstein, S.. Fischhoff, B., 1977. Do Those Who Know More Also Know More About How Much They Know? *Organizational Behavior and Human Performance, 20 (2)*, 159-183.

List, J.A., 2004. Neoclassical Theory Versus Prospect Theory: Evidence from the Marketplace. *Econometrica, 72 (2),* 615–625.

Mahajan, J., 1992. The Overconfidence Effect in Marketing Management Predictions. *Journal of Marketing Research, 29*, 329-342.

Malmendier, U., Tate, G., 2005. CEO Overconfidence and Corporate Investment. *Journal of Finance, 60 (6),* 2661–2700.

Meloy, M.G., 2000. Mood-Driven Distortion of Product Information. *Journal of Consumer Research, 27*, 345-359.

Meloy, M.G., Russo, E.J., and Miller, E.G., 2000. Monetary Incentives and Mood. *Journal of Marketing Research, 43,* 267–275.

Menkhoff, L., Schmidt, U., and Brozynski, T., 2006. The Impact of Experience on Tisk Taking, Overconfidence, and Herding of Fund Managers: Complementary Survey Evidence. *European Economic Review, 50 (7*), 1753-1766.

Milne, F., 2003. Finance Theory and Asset Pricing. Oxford University Press.

Moore, D.A., 2007. Not So above Average after All: When People Believe They Are Worse than Average and Its Implications for Theories of Bias in Social Comparison. *Organizational Behavior and Human Decision Processes, 102,* 42–58.

Moore, D.A., Cain, D.M., 2007. Overconfidence and Underconfidence: When and Why People Underestimate (and Overestimate) the Competition. *Organizational Behavior and Human Decision Processes, 103 (2)*, 197-213.

Moore, D.A., Healy, P.J., 2007. The Trouble With Overconfidence. *Psychological Review, 115 (2),* 502–517.

Moore, D.A., Oesch, J.M., Zietsma, C., 2007. What Competition? Myopic Self-Focus in Market-Entry Decisions. *Organization Science, 18 (3),* 440–454.

Murphy, A.H., Winkler, R.L., 1984. Probability Forecasting in Meteorology. *Journal of the American Statistical Association, 79 (387),* 489-500.

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

Nelson, M.W., Bloomfield, R., Hales, J.W., and Libby, R., 2001. The Effect of Information Strength and Weight on Behavior in Financial Markets. *Organizational Behavior and Human Decision Processes, 86 (2),* 168–196.

Nelson, M.W., Krische, S.D., and Bloomfield, R.J., 2003. Confidence and Inverstors' Reliance on Disciplined Trading Strategies. *Journal of Accounting Research, 41(3),* 503-523.

Noeth, M., Weber, M., 2003. Information Aggregation with Random Ordering: Cascades and Overconfidence. *The Economic Journal 113 (484),* 166–189.

Oenkal, D., Yates, F.J., Simga-Mugan, C., and Oetzin, S., 2003. Professional vs. Amateur Judgment Accuracy: The Case of Foreign Exchange Rates. *Organizational Behavior and Human Decision Processes 91,* 169–185.

Olsson, H., Juslin, P., 2000. The Sensory Sampling Model: Theoretical Developments and Empirical Findings. *Food Quality and Preference, 11,* 27-34.

Olsson, H., Winman, A., 1996. Underconfidence in Sensory Discrimination: The Interaction between Experimental Setting and Response Strategies. *Perception & Psychophysics, 58 (3),* 374-382.

Parikh, A., McReelis, K., and Hodges, B., 2001. Student Feedback in Problem Based Learning: A Survey of 103 Final Year Students Across Five Ontario Medical Schools. *Medical Education, 35 (7),* 632 - 636.

Petrusic, W.M., Baranski, J.V., 1997. Context, Feedback, and the Calibration and Resolution of Confidence in Perceptual Judgments. *American Journal of Psychology, 110 (4),* 543-572.

Pfeifer, P.E., 1994. Are We Overconfident in the Belief That Probability Forecasters Are Overconfident? *Organizational Behavior and Human Decision Processes, 58*, 203-213.

Plott, C.R., Zeiler, K., 2005. The Willingness to Pay–Willingness to Accept Gap, the "Endowment Effect," Subject Misconceptions, and Experimental Procedures for Eliciting Valuations. *American Economic Review, 95 (3),* 530-545.

Plott, C.R., Zeiler, K., 2007. Exchange Asymmetries Incorrectly Interpreted as Evidence of Endowment Effect Theory and Prospect Theory? *American Economic Review, 97 (4),* 1449-1466.

Price, P.C., 1998. Effects of a Relative-Frequency Elicitation Question on Likelihood Judgment Accuracy: The Case of External Correspondence. *Organizational Behavior and Human Decision Processes, 76 (3),* 277-297.

Pulford , B.D., Colman, A.M., 1997. Overconfidence: Feedback and Item Difficulty Effects. *Personality and Individual Differences, 23 (1),* 125-133.

Rapoport, A., Seale, D.A., and Winter, E., 2002. Coordination and Learning Behavior in Large Groups with Asymmetric Players. *Games and Economic Behavior, 39,* 111–136.

Rosenthal, R., Rosnow, R.L., 2006. Essentials of Behavioral Research: Methods and Data Analysis. 2nd edition, Academic Internet Publishers.

Schneider, S.L., 1995. Item Difficulty, Discrimination, and the Confidence-Frequency Effect in a Categorical Judgment Task. *Organizational Behavior and Human Decision Processes, 61 (2),* 148-167.

Shiller, R.J., 2003. From Efficient Markets Theory to Behavioral Finance (in Symposium: Financial Market Efficiency). *Journal of Economic Perspectives, 17 (1),* 83-104.

Shleifer, A., 2000. Inefficient Markets: An Introduction to Behavioral Finance. Oxford University Press.

Shore, T.H., Adams, J.S., Tashchian, A., 1998. Effects of Self-Appraisal Information, Appraisal Purpose, and Feedback Target on Performance Appraisal Ratings. *Journal of Business and Psychology, 12 (3),* 283-298.

Sieck, W.R., Merkle, E.C., Van Zandt, T., 2007. *Organizational Behavior and Human Decision Processes, 103,* 68–83.

Soll, J.B, 1996. Determinants of Overconfidence and Miscalibration: The Roles of Random Error and Ecological Structure. *Organizational Behavior and Human Decision Processes, 65 (2),* 117-137.

Stone, D.S., 1994. Overconfidence in Initial Self-Efficacy Judgments: Effects on Decision Processes and Performance. *Organizational Behavior and Human Decision Processes, 59 (3),* 452-474.

Stracca, L., 2004. Behavioral Finance and Asset Prices: Where Do We Stand? *Journal of Economic Psychology, 25,* 373-405.

Suantak, L., Bolger, F., and Ferrell, W.R., 1996. The Hard-Easy Effect in Subjective Probability Calibration. *Organizational Behavior and Human Decision Processes, 67 (2),* 201-221.

Subbotin, V., 1996. Outcome Feedback Effects on Under- and Overconfident Term Judgments (General Knowledge Tasks). *Organizational Behavior and Human Decision Processes, 66 (3),* 268-276.

Svenson, O., 1981. Are We All Less Risky and More Skillful Than Our Fellow Drivers? *Acta Psychologica, 47 (2),* 143-148.

Thomson, M.E., Oenkal-Atay, D., Pollock, A.C., and Macaulay, A., 2003. The Influence of Trend Strength on Directional Probabilistic Currency Predictions. *International Journal of Forecasting, 19,* 241–256.

Torngren, G., Montgomery, H., 2004. Worse Than Chance? Performance and Confidence Among Professionals and Laypeople in the Stock Market. *Journal of Behavioral Finance, (5), (3),* 148-153.

Tyszka, T., Zielonka, P., 2002. Expert Judgments: Financial Analysts Versus Weather Forecasters. *Journal of Psychology and Financial Markets, 3 (3),* 152–160.

Van Boven, L., Loewenstein, G., Dunning, D., 2003. *Journal of Economic Behavior and Organization, 51,* 351–365.

Van den Steen, E., 2004. Rational Overoptimism (and Other Biases). *American Economic Review, 94 (4),* 1141-1151.

Vigna, S.D., Mamendier, U., 2006. Paying Not to Go to the Gym. *American Economic Review, 96 (3),* 694-719.

Wallsten, T.S., Budescu, D.V., and Zwick, R., 1993. Comparing the Calibration and Coherence of Numerical and Verbal Probability Judgments. *Management Science, 39 (2),* 176-190.

Winman, A., 1997. The Importance of Item Selection in “Knew-It-All-Along” Studies of General Knowledge. *Scandinavian Journal of Psychology, 38,* 63–72.

Yaniv, I., Foster, D.P., 1995. Graininess of Judgment Under Uncertainty: An Accuracy-Informativeness Trade-Off. *Journal of Experimental Psychology: General, 124 (4),* 424-432.

Yaniv, I., Foster, D.P., 1997. Precision and Accuracy of Judgmental Estimation. *Journal of Behavioral Decision Making, 10,* 21-32.

Yates, J.F., Lee, J.W., and Bush, J.G.G., 1997. General Knowledge Overconfidence: Cross-National Variations, Response Style, and "Reality." *Organizational Behavior and Human Decision Processes, 70 (2),* 87-94.

3.6 Appendix

Summary of reviewed experiments

Legend:

Yes - yes
OK – more or less OK
No – no
? – cannot determine
RA – risk aversion
OC, UC – overconfidence, underconfidence
$ATP – monetary incentives paid according to performance
$flat – participation fee only

GKQ

GKQ	Mahajan (1992)	Frankenberger & Albaum (1997)
Representative stimuli	OK	?
Representation	Probabilities	OK
Subject pool	Students	Students
Incentives	In 1st experiment No In 2nd experiment top 1/3 got $10	Extra credit toward a course
Feedback	Yes (false)	No
Alternative explanations	Incentives, probabilities?	Incentives?
External/ecological validity	OK	OK
Result	OC, contradictory evidence ↗ performance ↘OC, more OC if expertise	OC – low-involvement, UC – high-involvement

GKQ	Hoelzl & Rustichini (2005)	Kovalchik et al. (2005)
Representative stimuli	Yes	?
Representation	Yes	OK
Subject pool	Students, recruitment problem	Students + elderly
Incentives	$ATP/No	?
Feedback	No	No
Alternative explanations	Recruitment strategy?	Incentives, representativeness?
External/ecological validity	OK	OK
Result	OC/UC if money, task difficulty	OC-students, elderly OC only on high confidence level

GKQ	Kogan (2006)	Wallsten et al. (1993)
Representative stimuli	OK	Yes
Representation	OK	OK
Subject pool	Students	Students
Incentives	$ATP	$ATP, $flat
Feedback	Yes	No
Alternative explanations	OC measured in different domain	Verbal scale?
External/ecological validity	OK	OK
Result	OC	OC, Verbal OC > numerical OC

Confidence intervals

CI	Klayman et al. (1999)	Juslin et al. (2003)
Representative stimuli	Yes	Yes
Representation	OK	OK
Subject pool	Students	Students
Incentives	$flat	$flat (+ some extra $?)
Feedback	No	No
Alternative explanations	Flat incentives?	Incentives?
External/ecological validity	OK	OK
Result	Big OC, OC differs among domains, not function of difficulty	Great OC, more in sensory than GK

CI	Cesarini et al. (2005)	Kirchler & Maciejovsky (2002)
Representative stimuli	Yes	OK – random dividend, info
Representation	OK	Probabilities
Subject pool	Students	Students (4/5 econ)
Incentives	$ATP	$ATP
Feedback	No	History, div. paid after each per.
Alternative explanations	Non-familiarity with CI	Students?
External/ecological validity	OK	? – students
Result	OC, repeating decreased OC	CI - OC, not generally OC in confidence of point estimates

CI	Biais et al. (2005)	Oenkal et al. (2003)
Representative stimuli	No - not random sample of questions	Yes
Representation	? – probabilities	OK
Subject pool	Students	Students + professionals
Incentives	Grade influence	No
Feedback	No – only realization, payoff	Yes
Alternative explanations	Students, RA, incentives, external measure of OC (other domain)	Incentives, non-familiarity with CI?
External/ecological validity	OK	OK
Result	CI – OC, OC ↘ trading performance	OC, more if CI

CI	Bolger & Oenkal-Atay (2004)	Du & Budescu (2007)
Representative stimuli	OK – constructed series, reflect reality	? – selected time series
Representation	? – graphs	? - graphs
Subject pool	Students	Students
Incentives	Extra credit in a course	$ lottery ATP
Feedback	Yes	No
Alternative explanations	Incentives, enough information?	Time series selection, graphs?
External/ecological validity	Information?	Yes
Result	OC ↘ over time (with feedback)	OC, UC depending on CI %

CI	Meloy et al. (2006)	Budescu & Du (2007)
Representative stimuli	?	Yes
Representation	OK	? - graphs
Subject pool	Students	Students
Incentives	$ATP or $flat	$ATP

Feedback	No	No?
Alternative explanations	? – Representativenes?	Graphs?
External/ecological validity	OK	OK
Result	OC	OC

Forecasting

Forecasting	Andersson et al. (2005)	Bolger & Harvey (1995)
Representative stimuli	Yes	Yes
Representation	OK	Probabilities
Subject pool	Students + experts	Students
Incentives	$ATP, no for experts	?
Feedback	No	No
Alternative explanations	Questionnaires, no incentives for experts	Incentives, probabilities?
External/ecological validity	Yes, for this type of prediction	?
Result	OC among experts	overestimation of p<0.5, underestimation of p>0.5

Forecast	Bolger & Oenkal-Atay (2004)	Kirchler & Maciejovsky (2002)
Representative stimuli	OK – constructed series, reflect reality	OK – random dividend, info
Representation	? – graphs	Probabilities
Subject pool	Students	Students (4/5 econ)
Incentives	Extra credit in a course	$ATP
Feedback	Yes	History, div. paid after each per.
Alternative explanations	Incentives, enough information?	Students?
External/ecological validity	Information?	? – students
Result	OC ↘ over time (with feedback)	CI - OC, not generally OC in confidence of point estimates

Forecasting	Oenkal et al. (2003)	Thomson et al. (2003)
Representative stimuli	Yes	simulated data?
Representation	OK	Graphs?
Subject pool	Students + professionals	Students + professionals
Incentives	No	No
Feedback	Yes	No
Alternative explanations	Incentives, non-familiarity with CI?	Incentives, out of laboratory, data?
External/ecological validity	OK	OK
Result	OC, more if CI	Hard-easy effect affects professionals more OC/UC than students

Forecasting	Torngren & Montgomery (2004)	Tyszka & Zielonka (2002)
Representative stimuli	Yes	Yes
Representation	?	OK
Subject pool	Students + professionals	Professionals
Incentives	No	No
Feedback	No	Yes
Alternative explanations	Incentives?	Incentives?
External/ecological validity	OK	OK
Result	Professionals worse performance than chance, more OC than stud.	More OC among fin. analysts than weather forecasters (better hit rate)

Forecast	Du & Budescu (2007)	Budescu & Du (2007)
Representative stimuli	? – selected time series	Yes
Representation	? - graphs	? - graphs
Subject pool	Students	Students
Incentives	$ lottery ATP	$ATP
Feedback	No	No?
Alternative explanations	Time series selection, graphs?	Graphs?
External/ecological validity	Yes	OK
Result	OC, UC depending on CI %	OC

Forecasting	Kelemen et al. (2007)
Representative stimuli	OK
Representation	OK
Subject pool	Students
Incentives	Course credit
Feedback	No
Alternative explanations	Incentives?
External/ecological validity	OK
Result	OC, decreased with practice

Market-entry games

Market-entry games	Camerer & Lovallo (1999)	Hogarth & Grieco (2004)
Representative stimuli	OK – not an issue here	No random sampling, no ref. class
Representation	OK – not an issue here	Probabilities
Subject pool	Students	Students
Incentives	$ATP	$ATP
Feedback	# of entrants only	# of correct answers
Alternative explanations	Not controlled for RA	Not controlled for RA
External/ecological validity	? – Fixed number of entrants	OK – no implications
Result	OC -> excess entry	No OC, no difficulty dependence, ambiguity

Market-entry games	Moore & Cain (2007)	Moore et al. (2007)
Representative stimuli	No random sampling, no ref. class	?
Representation	OK – not an issue here	OK
Subject pool	Students	Students
Incentives	$ATP	$ATP
Feedback	Full feedback	Yes, only once, only payoff
Alternative explanations	Not controlled for RA	Not controlled for RA
External/ecological validity	Subjects, questions?	students
Result	UC in difficult skill-based tasks, robust to experience, feedback, market forces	OC simple task, UC diff. task

Auctions

Auctions	Kirchler & Maciejovsky (2002)	Biais et al. (2005)
Representative stimuli	OK – random dividend, info	No - not random sample of questions
Representation	Probabilities	? – probabilities
Subject pool	Students (4/5 econ)	Students
Incentives	$ATP	Grade influence
Feedback	History, div. paid after each per.	No – only realization, payoff
Alternative explanations	Controlled for RA	RA not controlled
External/ecological validity	? – students	Students, RA, incentives, external measure OC (other domain)
Result	CI - OC, not generally OC in confidence of point estimates	CI – OC, OC ↘ trading performance

WTB/WTS

Willingness to sell/buy	Fellner et al. (2004)	Dittrich et al. (2005)
Representative stimuli	Yes	Yes
Representation	? – probabilities	? – probabilities
Subject pool	Students (mostly econ)	Students (mostly econ)
Incentives	$ATP	$ATP
Feedback	Yes – only payoff, played twice	No
Alternative explanations	Students + complicated task, mechanism	Students + complicated task, mechanism
External/ecological validity	? – students – illusion of expertise	? – students ?
Result	OC about own expertise	OC

Information

Information	Nelson et al. (2001)	Nelson et al. (2003)
Representative stimuli	OK	OK
Representation	OK – frequencies	OK
Subject pool	Students	Students
Incentives	$ATP	$ATP
Feedback	No	Yes
Alternative explanations	Not controlled for RA, signal creation well-understood?	Not controlled for RA, WTA/WTB disparity
External/ecological validity	OK	OK
Result	OC (UC) – high (low) strength, low (high) weight signal	OC, overreaction to information

Information	Noeth & Weber (2003)	Kraemer et al. (2006)
Representative stimuli	OK	OK
Representation	? – probabilities	? – probabilities
Subject pool	Students	Students
Incentives	$ATP	$ATP
Feedback	Yes	Yes
Alternative explanations	Probabilities?	Probabilities?
External/ecological validity	OK	OK
Result	OC about own signal	OC about the value of info

Assessment of others

Assessment of others	Shore et al. (1998)	Brenner et al. (1996)
Representative stimuli	?	OK
Representation	OK	OK
Subject pool	Students	Students
Incentives	No	$flat
Feedback	Yes	No
Alternative explanations	Incentives, representativeness of stimuli?	Incentives
External/ecological validity	OK	OK
Result	Higher self-assessment -> higher rating	OC

Assessment of others	Gershoff & Johar (2006)
Representative stimuli	? –choice of friends?
Representation	N.A.
Subject pool	Students + ?
Incentives	No
Feedback	Yes
Alternative explanations	Incentives, the choice of subjects
External/ecological validity	OK
Result	OC

Self-awareness questions

Self-awareness	Ehrlinger et al. (2008)	Niederle & Vesterlund (2007)
Representative stimuli	No	Yes
Representation	Yes	Yes
Subject pool	Students, non-students	Students
Incentives	Extra course credit, $5	$ATP
Feedback	No	Yes
Alternative explanations	Representative stimuli, Insufficient information?	No
External/ecological validity	OK	OK
Result	OC-unskilled, UC-skilled	Men OC

Self-awareness	Hoelzl & Rustichini (2005)
Representative stimuli	Yes
Representation	Yes
Subject pool	Students, recruitment problem
Incentives	$ATP/No
Feedback	No
Alternative explanations	Recruitment strategy?
External/ecological validity	OK
Result	OC/UC if money, task difficulty