Development of the overconfidence measurement instrument for the economic experiment

Michailova, Julija

Christian Albrechts University of Kiel

2010

DEVELOPMENT OF THE OVERCONFIDENCE MEASUREMENT INSTRUMENT FOR THE ECONOMIC EXPERIMENT

Dr. J. Michailova

Quantitative Economics, CAU[1]
Neufeldtstrasse 10
24118 Kiel, Germany
Telephone: +491708444578

E-mail: julija_michailova@yahoo.com

2nd November, 2011.

Abstract

In this article results of the two experiments, aimed at the development of the instrument (test) that would enable construction of the comprehensive measure of individual overconfidence for the use in economic overconfidence experiments, are presented. Instrument was obtained in a two-stage procedure. In the first experimental phase, a pilot test, consisting of fifty general-knowledge questions of the unknown difficulty, was conducted to divide the items into three difficulty levels: hard, average-difficulty and easy questions. The second phase was aimed at verification of the replicability of results. Statistical tests supported the existence of the hard-easy effect, verified the success of categorization of questions into three levels of difficulty, and showed that gender was not associated with overconfidence in the developed instrument. The average group overconfidence measures obtained from both experimental phases did not differ from each other significantly. Instrument's internal consistency was found to be good and acceptable for the use in social research. Compared to the tests used in the foregoing economic experiments, the obtained test is believed to result in the improvement of the overconfidence measurement quality.

Keywords: overconfidence, instrument development, experiment.

JEL codes: C42, C43, C80, C99

Acknowledgements: Author wants to express gratitude to Dr. Briony D. Pulford for a fruitful cooperation. I thank Prof. Dr. Jürgen Golz for valuable comments. A special thank to my colleague Christian Radden, who helped in translating the experimental instructions to German. I acknowledge a German Academic Exchange Office (DAAD) scholarship.

[1] Present address: Helmut-Schmidt University, Chair of Political Economy and Empirical Research in Economics, Holstenhofweg 85, Hamburg, Germany.

1 INTRODUCTION

A large body of economic literature presents results of experiments on overconfidence. The concept of overconfidence is based on the evidence form cognitive psychological research, which suggests that human-beings overestimate their knowledge, abilities and precision of their information. As example Bar-Hillel (2001) points out that, when subjects are P% sure that they have answered a question correctly in fact they are right on average less than P% of the time. There is plenty of evidence for people to be in general overconfident, and phenomenon of overconfidence has been found in many different samples of the population, e.g. students (Fischhoff, Slovic and Lichtenstein,1977; Koriat et al., 1980; Zakay and Glicksohn, 1992), members of the armed forces (Hazard and Peterson, 1973), CIA analysts (Cambridge and Shreckengost, 1978), entrepreneurs (Baron, 2000), clinical psychologists (Oskamp, 1962), bankers (Staël von Holstein, 1972), executives (Moore, 1977), negotiators (Neale and Bazerman, 1990), managers (Russo and Schoemaker,1992), lawyers (Wagenaar and Keren, 1986), and civil engineers (Hynes and Vanmarcke, 1976). Overconfidence is already present in children (see Powel and Bolich, 1993; Allwood, Granhag, and Jonsson, 2006), and boys are found to be more overconfident than girls (e.g. Sieber, 1979; Newman, 1984; Allwood et al., 2006). However, in adult samples no differences between both genders in overconfidence[2] are observed (e.g. Lichtenstein and Fischhoff, 1981; Gigerenzer et al., 1991).

In economic experiments, there is no conventional method of measurement of the inborn level of subjects' overconfidence. For this purpose various proxies, tests and tasks are used, that not always offer a satisfactory measure of individual overconfidence. The need for this research is stimulated by the fact that previous experiments have drawbacks in the way they measure overconfidence, and thus overconfidence might have been caused (to some extent) by other reasons than the imperfection of human nature, but rather by the mistakes in the tests'/ tasks' construction. Findings from psychological research indicate that the observable biases in judgment are often result of the inappropriateness of the task, e.g. a task is unclear to subjects, one gender finds task more difficult than the other, or there is not enough motivation for active participation. Thus development of the overconfidence test was implemented with the following assumptions in mind. First of all, most of the foregoing researchers followed the famous work by Russo and Schoemaker (1992) and used interval elicitation tasks to assess overconfidence. However, these tasks are prone to produce extreme overconfidence (see

[2] Assessed via sets of general knowledge questions.

Klayman et al., 1999). Second, previous authors used tests that were neither balanced to the hard-easy effect nor country or gender balanced. Yet, unbalanced tests can artificially create high levels of under- or overconfidence either in the whole group, or in parts of it. Third, overconfidence was often assessed based on the insufficient number of assignments or test items; psychological studies of overconfidence use amounts of items that are much higher. And last but not least, many of the tasks and tests were either not administered, or were not (financially) rewarded.

This paper presents the results of the two experiments, aimed at the development of the instrument (test) that would enable construction of a comprehensive measure of individual overconfidence for the use in economic overconfidence experiments dealing with: 1) the role of overconfidence in occurrence of stock-prices' bubbles, and 2) impact of overconfidence and risk aversion on economic behavior of individual traders. Test is intended for the detection of potential experimental subjects with high and low degrees of overconfidence and their subsequent grouping into two types of asset markets: rational and overconfident. Hence, a well-designed instrument should allow assessment of differences between the subjects with respect to their overconfidence and minimize the measurement error.

The developed test differs from those used in prior economic experiments in some important respects. First, another test format was chosen, namely multiple choice discrete propositions' task format, which is clearer to subjects and is not inherently prone to production of extreme overconfidence levels. Second, test was balanced to the hard-easy effect, by the inclusion of an equal number of questions of three difficulty levels (hard, medium-difficulty and easy). Third, in construction of the test it was controlled for the possible country and gender bias, e.g. no inclusion of questions that might be easier to one gender. And finally, compared to some studies, the test is expanded to include more items. Instrument was obtained in a two-stage procedure in which a pilot test was used to assess questions' difficulty, based on the group accuracy in answering every item of the 50 initial. Then six questions of each of the three difficulty types were chosen for the inclusion in the final test. The second experimental phase was aimed at verification of replicability of results, namely of the average degree of group overconfidence, the obtained categorization into three difficultly levels and of controlling for the gender bias. Both experiments were conducted with the students enrolled into different disciplines of social sciences. Experimental sessions were administered and subjects were offered a reward, on the basis of competition in test accuracy. The final instrument consists of 18 general knowledge questions unrelated to economics, financial markets or experiments. Questions are not connected to economics, as otherwise they could

cause biased results if the same test is used with a heterogeneous pool of subjects[3]. Evidence was found for the significant effect of the question difficulty on the overconfidence measure and existence of the gender bias. Compared to medium and easy questions, which resulted in under-confidence, hard questions produced significantly higher levels of overconfidence. The three types of questions also significantly varied from each other in terms of the produced confidence and accuracy. This result verified the success of categorization of questions into 3 levels of difficulty in the created overconfidence measurement test. In the initial instrument as much as 16% of variance in accuracy and 7% of variance in confidence was explained by gender. In the final test gender is not associated with overconfidence, and there is almost no variance in confidence and accuracy that is gender dependent.

Paper proceeds as follows. In Section 2 a review of the findings of psychological overconfidence literature are presented. In Section 3 findings from the theoretical and empirical research on overconfidence in finance are introduced, and ideas, on how overconfidence was measured in the previous experimental research, are presented; in closing of this Section a problem statement with the ideas about research improvement are provided. In Section 4 methodology of the test construction is described. In Section 5 statistical data analysis is presented. In Section 6 findings from the experiment with the final overconfidence measurement instrument are analyzed, and, finally Section 7 concludes.

2 OVERCONFIDENCE IN PSYCHOLOGICAL RESEARCH

Our life is full of uncertainty, and many decisions are based on beliefs concerning the likelihood of uncertain events (Tversky and Kahneman, 1982). These beliefs can be expressed in numerical form as subjective probabilities. Question of generation of these probabilities is one of the most important topics in the area of cognitive psychology (Bar-Hillel, 2001). Bar-Hillel (2001) suggests that subjective probabilities are not just imperfect or inaccurate versions of objective probabilities, but rather are governed by cognitive principles of their own. To generate subjective probabilities, people rely on a limited number of heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations. In general, these heuristics are quite useful, but sometimes they lead to severe and systematic errors (Tversky and Kahneman, 1982), and thus to non-optimal judgments (see Russo and Schoemaker, 1992, for the description of the heuristics employed by people for assessment of probabilities). Use of heuristics for generation of

[3] Deaves et al. (2004) also motivate their choice of non-economic questions by the attempt "to avoid giving either group of participants a relative advantage because of subject content".

subjective probabilities is a cognitive cause of overconfidence. Overconfidence is characterized by the tendency to overestimate one's skills, prospects for success, the probability of positive outcomes or the accuracy of one's knowledge, and arises from not knowing the limits of one's knowledge (Conger and Wolstein, 2004).

2.1 Definition of the Overconfidence Bias

In financial literature there are several findings that are often summarized under the concept of overconfidence: miscalibration, the better than average effect, illusion of control, and unrealistic optimism. The issue of whether these notions are related is mainly unexplored[4] (Glaser and Weber, 2007). In psychological research, however, only miscalibration is defined as overconfidence.

Miscalibration

Miscalibration is a cognitive bias that rests on the fact that people tend to overestimate the precision of their knowledge. In the experiments on calibration, participants answer a series of (general knowledge) questions and stipulate their confidence of being correct for each answer. Calibration is tested by comparing the percentage of questions that a participant has answered correctly with the participant's average confidence in the answers to these questions. Individuals are considered to be well calibrated if the following condition is satisfied: over the long run of those responses made with confidence P, about P% should be correct (Adams, 1957). However most of the people are not well-calibrated and demonstrate overconfidence (miscalibration), which manifests itself through a systematic deviation from perfect calibration and is defined as an "unwarranted belief in the correctness of one's answer" (Lichtenstein, Fischhoff, and Phillips, 1977). Typically, for all questions the proportion of correct answers is lower than the assigned probability (Lichtenstein, Fischhoff, and Phillips, 1982). In a second strand of psychological literature, where overconfidence is measured by asking subjects to state for a series of questions with unknown numerical answer an upper and lower limit such that a subject is X% sure that the real answer would fall into that interval, the usual finding is that subjects' probability distributions are too tight (Lichtenstein, Fischhoff, and Phillips, 1982). E.g. when subjects are asked to state 90% confidence intervals for some uncertain quantity, the percentage of true values that fall outside the interval, is higher than 10% (the percentage of surprises of a perfectly calibrated person). In the study of Alpert and

[4] Oberlechner and Osler (2003), Glaser, Langer, and Weber (2005), Glaser and Weber (2007) found no significant correlation between miscalibration and better than average effect measures.

Raiffa (1982) 50% intervals included the true quantity only about 30% of the time; 98% intervals included the true quantity only about 60% of the time.

Better than average effect

Inclination of people to exaggeration of their talents embodies itself in the better than average effect. Taylor and Brown (1988) document in their survey that people have unrealistically positive views of the self, i.e. they think about themselves as possessing above the average abilities (e.g. with regard to skills or positive personal traits) compared to other people. One of the most cited works by Svenson (1981) states that 82 percent of a group of students rank themselves among the 30 percent of drivers with the highest driving safety. Sümer et al. (2006) also found better than average effect among drivers in their sample, resulting from exaggerated ratings of self-reported driving skills.

Illusion of control and unrealistic optimism

Illusion of control is linked to the exaggeration of the degree to which one can control one's fate. Subjects prone to the illusion of control, tend to underestimate the role of chance in human affairs and to misperceive games of chance as games of skill (Kahneman and Riepe, 1998). Langer (1975) in her pioneering work defines this phenomenon as "an expectancy of a personal success probability inappropriately higher than the objective probability would warrant". The phenomenon of unrealistic optimism about future life events is a cognitive bias that is strongly related to the illusion of control (Weinstein, 1980). Johnson, McDermott, Barrett, Cowden, Wrangham, McIntyre, and Rosen (2006) point out that numerous empirical findings confirm "that mentally healthy people tend to exhibit psychological biases that encourage optimism, collectively known as "positive illusions". According to Kahneman and Riepe (1998) "most people's beliefs are biased in the direction of optimism". Griffin and Brenner (2005) note, that "optimistic overconfidence" represents overestimation of the probabilities of the events that are advantageous to the subject. Probabilities of the unfavorable events are underestimated by optimists; even in cases when they have no control over them, e.g. Kahneman and Riepe (1998) note "most undergraduates believe that they are less likely than their roommates to develop cancer or to have a heart attack before the age of fifty".

2.2 MEASUREMENT OF OVERCONFIDENCE

There are two types of calibration assessment techniques used in the psychological experiments: making probability judgments about discrete proposition*s*, and the calibration of probability density functions assessed for uncertain numerical quantities (the fractile method).

Measurement of Calibration with Discrete Propositions' Task

To measure overconfidence with the discrete propositions, subjects are suggested to answer a series of questions and state their confidence for each question that their answer was correct. Discrete propositions can give no alternatives for an answer, or suggest one, two, or multiple answer choices.

Calibration can be expressed through several various measures (e.g. Calibration curve, Brier score). However a convenient measure, enabling discrimination between under- and overconfidence, is the bias score. The bias score is calculated as the difference between the average confidence level across all questions and the proportion of correct answers. A positive bias score represents overconfidence, and a negative bias score represents underconfidence. A bias score of zero indicates an accurately calibrated (neutral) person.

$$bias\ score = average\ \%\ confidence - average\ \%\ correct \tag{1}$$

Or as in Pulford (1996):

$$over/underconfidence = \frac{1}{N}\sum_{t=1}^{T} n_t(r_t - c_t) \tag{2}$$

Here, T is the total number of response categories used, n_t is the number of times the response r_t was used and c_t is the proportion correct for all items assigned probability r_t.

Measurement of Calibration with the Fractile Method

Lichtenstein, Fischhoff, and Phillips (1982) note that uncertainty about the value of an unknown continuous quantity (e.g. what is the air distance from London to Tokyo) "may be expressed as probability density function across the possible values of that quantity". The assessor has to state values of the uncertain quantity that are associated with a small number of predetermined fractiles (quantiles) of the distribution, i.e., as mentioned before, to state, for a series of questions with unknown numerical answer, upper and lower limits such that she is X% sure that the real answer would fall into that interval. There are two calibration measures for continuous items: interquartile index and surprise index. Interquartile index is the percentage of items for which the true value falls inside the interquartile range (i.e., between the 0.25 and 0.75 fractiles) (Lichtenstein et al., 1982). Interquartile index of a perfectly calibrated person is 0.5. Surprise index is the percentage of true values that falls outside the most extreme fractiles assessed. Lichtenstein et al. (1982) write that when the most extreme fractiles are assessed as 0.01 and 0.99, the surprise index of the perfectly calibrated person should be 2. Large surprise index shows the inability of the assessor to state confidence

bounds wide enough to include as much as possible of the true values. This indicates overconfidence.

2.2 Factors Influencing the Degree of Overconfidence

Keasey and Watson (1989) identified four factors that have an impact on the accuracy-confidence relationship: task complexity, amount of feedback, subjects' level of motivation, and their skills.

Hard-Easy Effect

The degree of overconfidence is connected to the complexity of the task. This is called the "had-easy" effect. This effect occurs when the degree of overconfidence increases with the increase in the difficulty of the questions, where difficulty is measured as the percentage of correct answers (Gigerenzer, Hoffrage, and Kleinbölting, 1991). Lichtenstein et al. (1982) suggest that "the most pervasive finding in recent research is that people are overconfident with general-knowledge items of moderate or extreme difficulty". Many studies have supported this conclusion, e.g. Fischhoff, Slovic and Lichtenstein (1977), Koriat, Lichtenstein and Fischhoff (1980), Nickerson and McGoldrick (1965), Cambridge and Shreckengost (1978), Arkes, Christensen, Lai, and Blumer, (1987), Ronis and Yates, (1987), Sniezek, Paese, and Switzer, (1990) and etc. The degree of overconfidence is the highest with the tasks of high difficulty (e.g. Clarke, 1960; and Pitz, 1974); as tasks get easier, overconfidence is reduced (Lichtenstein et al., 1982).

Motivation and Feedback

The two ways of increasing subjects' calibration are: motivation through reward for their assessment to be more precise, and outcome feedback (Lichtenstein, Fischhoff, and Phillips, 1982).

Motivation, according to Bohner et al. (1998) is one of the factors that encourage people to abandon the use of effort minimizing heuristics in favor of more effortful probabilities' estimation strategies; thus it has an impact on the accuracy-confidence relationship. Motivation through reward is named to be a tool, helping to improve subjects' calibration, by Lichtenstein et al. (1982). This finding is supported by the paper of Hoelzl et al. (2005) who have discovered a significant change in overconfidence pattern depending on the existence or nonexistence of the monetary reward.

Lichtenstein, Fischhoff, and Phillips (1982) state, that receiving outcome feedback after every assessment is the best condition for successful training to be better calibrated. Adams and

Adams (1958) have found modest improvement in subjects' calibration after five training sessions where they were given feedback on their performance. Fischfoff (1982) reports some successful training exercises, mostly using large amounts of well-structured feedback. Lichtenstein and Fischhoff (1980) studied the impact of extensive, personalized calibration feedback on two groups of subjects. Perceptible improvement in calibration was reported, however no improvement was found in probabilities' assessment by fractile method. In general, improvement in the accuracy of estimates is difficult to achieve (see e.g. Ferrell and McGoey, 1980; Lichtenstein and Fischhoff, 1980; Koriat et al., 1980), and there are reasons to be pessimistic about how well training transfers across time or tasks (Camerer, 1995).

3 OVERCONFIDENCE IN FINANCIAL LITERATURE

3.1 FINDINGS FROM THE RESEARCH

Following psychological research in overconfidence, interest in the consequences of economic subjects' overconfidence on financial decision making, functioning of markets and economic outcomes has occurred in behavioral economics. Findings of behavioral finance have an important value in understanding various anomalies and stylized facts found for example in the stock market. Overconfidence research in economics is developing in two directions: theoretical modeling and empirical testing of these models.

Theoretical models about the impact of overconfidence on the processes in financial markets and the behavior of investors are based on the initial assumption of traders' overconfidence, whose decision-making is modeled according to this premise. Behavioral finance models predict that overconfidence causes excess trading volume (De Bondt and Thaler, 1984; Shiller, 2000; Benos, 1998; Caballé and Sákovics, 2003), and excess price volatility (Scheinkman and Xiong, 2003; Benos, 1998, Daniel et al., 1998); it induces occurrence of the speculative price bubbles (Scheinkman and Xiong, 2003) and increases the market depth (Odean, 1999; Kyle and Wang, 1997; Benos, 1998); it makes markets underreact to abstract, statistical, and highly relevant information and overreact to salient, but less relevant information (Odean, 1998); it makes returns of financial assets predictable (Daniel et al., 1998, 2001; Scheinkman and Xiong, 2003). Overconfidence increases investors' tendency to herd (Hirshleifer, Subrahmanyam and Titman, 1994) and makes them choose riskier and undiversified portfolios (Odean, 1998, 1999; Lakonishok, Shleifer and Vishny, 1992), overconfident investors trade more aggressively, i.e. their trading activity is too high (Odean, 1999; Gervais and Odean, 2001) and their expected utility is reduced (De Long et al., 1991; Odean, 1998).

There are not so many empirical and experimental studies testing the assumptions of the abovementioned theories of the impact of overconfidence on financial decision making, functioning of the markets and economic outcomes. Empirical findings support the premise of theoretical models, that overconfidence results in high trading volume in the market (Statman, Thorley, and Vorkink, 2006; Kim and Nofsinger, 2003); it also increases the probability of bubbles' occurrence[5] (Oechssler, Schmidt and Schnedler, 2007). A higher degree of overconfidence reduces traders' performance/ welfare (Fenton-O'Creevy et al., 2003; Biais, Hilton, Mazurier, and Pouget, 2005; Odean, 1999; Barber and Odean, 2002; Nöth and Weber, 2003), and causes mistakes in financial decision making (Biais et al., 2005); unrealistically positive self-evaluation increases trading volume (Glaser and Weber, 2007). There is no clear conclusion about how overconfidence may influence markets' reaction to new information: e.g. Loughran and Ritter (1995) found that markets overreact to new information, and studies by Bernard and Thomas (1989, 1990), Michaely, Thaler and Womack (1995) detected underreaction. There is no clear relationship between the degree of overconfidence, and the degree of professionalism: in the studies by Kirchler and Maciejovsky (2002) and Glaser and Weber (2007) overconfidence increases with experience, whereas studies of Menkhoff, Schmeling, and Schmidt (2006) and Biais et al. (2005) find the reverse dependence.

3.2 Measurement of Overconfidence in Empirical and Experimental Studies

As mentioned above, there are a few empirical and experimental studies designed to test the impact of overconfidence on financial decisions, market outcomes and subjects' performance. Some of them present only an indirect evidence of such impact, as they measure overconfidence via various proxies and it is not always clear who of the subjects and how strong are overconfident. Other studies measure the inborn level of subjects' overconfidence via the different tasks and tests, related or non-related to economics and finance. Such tests usually enable construction of the overconfidence measure for each individual. Most often these tasks are related to confidence intervals' estimations in the spirit of the work by Russo and Shoemaker (1992).

Proxies for Overconfidence

Papers that use proxies for overconfidence do not allow for the numerical measurement of the degree of overconfidence. E.g. Statman, Thorley, and Vorkink (2006) test the hypothesis of

[5] Top-rank belief variable "has a positive and significant effect on the probability of bubbles" (Oechssler et al., 2007).

interdependence of overconfidence and high trading volume for the USA stock market. As a proxy for the degree of overconfidence authors suggest using high past returns. They argue that after high past returns posterior volume of trade will be higher, as successful investment increases the degree of overconfidence. The same proxy for overconfidence (i.e. high past returns) was utilized by Kim and Nofsinger (2003) for the Japanese stock market.

Barber and Odean (2001) use gender of the trader as a proxy for overconfidence. Their assumption is that, based on the psychological literature, women are less overconfident than men, thus they are going to trade less than men. In their study men are actually found to trade more than women. In another paper Barber and Odean (2002) employ as a proxy of overconfidence changes in the trading patterns and performance of the 1607 investors who switched from the phone-based trading to online trading between 1992 and 1995. They present evidence that these investors traded more actively and speculatively, and performed subpar.

Blavatsky (2008) measures overconfidence by the taken choice in a simple task: subjects can either bet on their knowledge or on the equivalent lottery. Those who choose an option to bet on their own knowledge are classified to be overconfident (others are underconfident). Under this measurement procedure, subjects, on average, exhibit underconfidence about their own knowledge, and their confidence does not depend on their attitude towards risk/ ambiguity.

Overconfidence Measured via Tests and Tasks

In comparison to the studies that use various proxies to measure overconfidence, questionnaire studies enable direct assessment of each subject's under- or overconfidence.

Kirchler and Maciejovsky (2002) investigate overconfidence within the context of an experimental asset market. Miscalibration of subjects is measured before each trading period, with the help of the two price prediction tasks: point prediction with the confidence in forecast, and 98% confidence interval prediction. Results presented in their paper indicate that in some periods participants demonstrate overconfidence and in others underconfidence, thus they are not generally prone to overconfidence. Kirchler and Maciejovsky (2002) also show that higher degree of overconfidence is negatively correlated with the earnings of the participants of their experiment.

Fenton-O'creevy, Nicholson, Soane, and Willman (2003) examine the impact of illusion of control on the performance of traders in four investment banks. They use a computer-based task measurement of the illusion of control to execute measurement of overconfidence: 107 participants had to raise an index on the computer screen by pressing keyboard-buttons and

rate their success in doing so from 1 (not at all successful) to 100 (very successful). The index in reality was modeled as random walk process with an upward trend, and thus the button had no influence on its development.

Biais, Hilton, Mazurier, and Pouget (2005) conduct an experiment to check if overconfidence has impact on subjects' trading performance (trading activity and profits). They use the scale adapted from Russo and Schoemaker (1992) to measure the degree of overconfidence in a group of 245 students. Their test consisted of 10 general-knowledge questions with known numerical answers for which subjects had to state 90% confidence intervals. Several weeks later, after the students' overconfidence was measured, they participated in the experimental asset market. Questions that were used to measure subject's miscalibration had nothing to do with financial markets, yet they affected strategies and performance in the experimental market; this points at the robustness of the psychological construct independent of the context in which the questions are asked (Biais et al., 2005).

In their stock market experiment Deaves, Lüders, and Luo (2004) are testing for premises that overconfidence leads to an increase in trading activity, and that gender influences trading activity through differences in overconfidence. They measure overconfidence of their subjects using a calibration based approach prior to conducting the experiment. Compared to the tests used in the other studies, their test contains more (up to 20) items. Each of the general knowledge questions in their test had a known numerical answer for which subjects had to state upper and lower bounds of 90% confidence interval in which the real answer would fall. Their choice of the non-economic questions is motivated by the attempt to avoid giving either group of participants a relative advantage because of subject content (Deaves et al., 2009).

Stotz and von Nitzsch (2005) in their paper investigate the extent of analyst overconfidence in their abilities to forecast prices and earnings. 112 bank analysts had to answer two questions: one asked them to rank their skills with regard to their price or earnings estimates in comparison to their colleagues, and another asked to estimate what percentage of analysts produce work superior/ inferior to them. Two types of coefficients, measuring subjects' overconfidence, were then calculated: overconfidence coefficients for earnings and in price targets. Results presented in their paper suggest, that overconfidence increases with an increasing perception of control.

Glaser and Weber (2007) asked a sample of approximately 3000 individual investors with online broker accounts to answer an online test, which enabled the authors to measure several manifestations of overconfidence: miscalibration, better than average effect, illusion of control, and unrealistic optimism. To measure miscalibration they asked subjects to state

upper and lower bounds of 90% confidence interval to the five economy-related questions[6] and five stock price predictions. Only 114 investors answered all their economy-related questions, and 165 – stock price prediction questions. By correlation the obtained measures of overconfidence and trading volume Glaser and Weber (2007) explored the connection between them.

Glaser, Langer, and Weber (2005) surveyed 123 professional traders and investment bankers, and compared results to a student control group in order to analyze whether professionals are prone to judgmental biases to the same degree as lay men. They measured overconfidence of their subjects by the means of four tasks: 1) subjects stated 90% confidence intervals for 20 knowledge questions (ten general knowledge and ten economics and finance knowledge); 2) subjects had to assess their performance in the knowledge task (how many right answers?) and assess own performance compared to others (how many right answers compared to the others?); 3) make 15 stock market forecasts by stating 90% confidence intervals, and 4) predict a trend in stock prices forecasting via confidence intervals. In most tasks the degree of overconfidence of professionals was significantly higher than of the student group.

To analyze the effect of professionalism on investment decisions Menkhoff, Schmeling, and Schmidt (2006) conducted a survey of approximately 500 subjects, consisting of professionals and lay men. Alongside with other aspects in their survey, they measured overconfidence via two questions on the "appropriate self-evaluation" (in other words better than average effect) in which subjects had to estimate their performance and information compared to the other investors. They find that, among other control variables, portfolio turnover is related to lower risk aversion and higher overconfidence.

Menkhoff, Schmidt and Brozynski (2006) surveyed 117 fund managers in order to detect an impact of experience on overconfidence, risk taking, and herding behavior. Their survey measures overconfidence via three tasks that enable assessment of the three manifestations of overconfidence: 1) evaluation of the own performance compared to the other fund managers (better than average effect), 2) 90% confidence estimation of the DAX index forecast (miscalibration), and 3) a third task is aimed at measurement subjects' illusion of control (subjects are asked to rate the statement: economic news are not surprising to me). They find that experienced fund managers tend to exhibit herding behavior to a lesser extent than inexperienced ones; while evidence concerning the impact of experience on risk taking and

[6] However they call them "general knowledge" questions.

overconfidence is mixed: positive-self evaluation and illusion of control are increasing with experience, whereas miscalibration on the contrary decreases.

In their paper Oechssler, Schmidt, and Schnedler (2007) study whether bubbles can occur in the experimental markets that pay no dividends on assets. To measure overconfidence in their experiment they asked subjects, prior to each round, to rank themselves among the 60 subjects of a treatment in terms of payoff of that round. For each period the percentage of subjects who ranked themselves to be better than median (rank 30 or higher) was compared to the expected number of 50%. Overconfidence in their experiment was modest as merely 54% of the subjects thought to be better than the median. They have also constructed a second variable – "top-rank belief" – that measured, for each round, the number of subjects who thought they would be the best in terms of payoff. This construct has positive and significant effect on the probability of bubbles' occurrence in their experimental market.

3.3 PROBLEM STATEMENT

This paper is aimed at the development of the instrument (test) that would enable construction of the comprehensive measure of individual overconfidence for the use in two economic experiments dealing with: 1) the role of overconfidence in occurrence of stock-prices' bubbles, and 2) impact of overconfidence and risk aversion on economic behavior of individual traders. A well-designed instrument will allow assessing differences of the subjects with respect to their overconfidence and minimize the measurement error.

In my opinion previous works have drawbacks in the way they measure overconfidence, and thus in prior experiments overconfidence might have been caused (to some extent) by other reasons than the imperfection of human nature, i.e. by the mistakes in the tests'/ tasks' construction[7]. Thus development of the overconfidence test was conducted with the following assumptions in mind:

From the review above one can see that overconfidence in financial settings is estimated either with the help of some assignments (e.g. estimate what percentage of analysts produce work superior to you?) or by the means of interval elicitation tests. However, overconfidence is often assessed based on the insufficient number of assignments or test items. Thus it raises doubts that these instruments actually offer a comprehensive measure of overconfidence of an individual. This fact is mentioned in the work of Menkhoff et al. (2006), who measured overconfidence with three assignments; Barber and Odean (2002) use only two assignments.

[7] Not to mention the studies in which overconfidence was never measured directly. For a review see Glaser and Weber (2007).

In comparison, the psychological studies of overconfidence use the amount of items that is much higher, and the minimum number of items for a reliable test is ten (Kline, 1993).

Most of the foregoing researchers followed the famous work by Russo and Schoemaker (1992) and used interval elicitation tests to assess overconfidence. However, interval estimation tasks are prone to produce extreme overconfidence (see e.g. Klayman et al., 1999). One reason to that is that subjects do not really understand the nature of these intervals and "what they are being asked to come up with" (Deaves et al, 2004). Also use of these instruments to measure the improvements in calibration, when the test is conducted before and after the experiment, is useless as this method does not allow for the improvement in calibration after training sessions; on the other hand subjective probability elicitation for the discrete items, combined with financial reward, can be improved (Lichtenstein and Fischhoff, 1980).

Findings from psychological research show that overconfidence is the most pronounced for the hard questions (few people know the right answer) and the least pronounced for the easy ones (most of the people know the correct answer). However, the abovementioned papers did not make use of the balanced to hard-easy effect tests. This could have artificially created high levels of under- or overconfidence, e.g. in the experiment of Deaves et al. (2009) none of the subjects got close to the perfect calibration measure, and even the best calibrated participants exhibited rather high degrees of overconfidence.

Connected to the hard-easy effect are country and gender biases. Country bias rests on the fact, that some questions might be easy in one country, but in another one they might be hard. Gender bias is produced by the choice of questions for the test that could be easier for men than women (e.g. sports, masculine hobbies) and vice versa. This could result in the inappropriate levels of under- or overconfidence for one gender compared to the other, or in one country compared to the other. Nevertheless, previous authors used tests that were not country or gender balanced, e.g. Deaves et al. (2009), used the same test in several locations. Finally, many of the tasks and tests discussed above were either not administered (e.g. Glaser and Weber (2007) conducted their survey via internet, and subjects might have used other sources than their own knowledge for answering the test), or were not (financially) rewarded.

Based on the abovementioned analysis, the developed instrument for measurement of overconfidence in the planned stock market experiments will differ in some important respects. First, another test format is chosen, namely multiple choice discrete propositions' task format, which, due to its simplicity, is clearer to subjects and not inherently prone to production of extreme overconfidence levels. Second, a pilot test is conducted to assess questions' difficulty and to single out easy, medium and hard questions. Then an equal

number of questions of the three difficulty levels are included in the final test. Third, in construction of the test it is controlled for possible country and gender biases, e.g. I have tried to avoid questions that might be easier to one gender than the other. Forth, to check if the categorization into three difficultly levels and controlling for gender bias is successful, final instrument is pre-tested with the target group of students, namely those who are enrolled in different disciplines of the social sciences. Fifth, overconfidence measurement phase of the experiment is administered and financially rewarded. Moreover it is rewarded on the competition in the test accuracy basis, which should discourage sharing the results among students and thus increase the reliability of the measurement. And finally, compared to some of the authors, my test is expanded to include more questions.

4 METHOD

Procedure and Subjects

A pilot test, whose purpose was to select questions for the final questionnaire, was conducted on the 19th May, 2008 at Christian-Albrechts University of Kiel. Subjects were given approximately 30 minutes time to fill in the 50 questions test at the end of the lecture on Social Politics. Three monetary prizes were offered for those participants who got the most questions right. A reward on the basis of competition in test accuracy was chosen in order to decrease the desire of subjects to share answers, and thus increase reliability of the obtained individual bias scores. Only fully completed tests were considered for the prize. A total of 96 tests were completed, of them 44 by males, and 52 by females. Most of the students were German (91 subjects). After the initial analysis 12 partially incomplete tests were not included in the further analysis. From the remaining 84 tests 50 were chosen randomly – 25 of men, and 25 of women. Participants of the test aged from 20 to 29 years (M = 24.32, SD = 0.31), and have studied from 3 to 11 semesters (M = 6.98, SD = 2.11). All participants were students of social sciences; of them 40% studied management, 38% were economics students, and 22% subjects were enrolled into other social studies. Average age of male subjects was 24.48 years (SD = 2.43), and their average duration of study was 7 semesters (SD = 2.27). Average age of female subjects was 24.16 years (SD = 1.97), and their average duration of study was 6.96 semesters (SD = 1.99). For information about participants' age and duration of studies refer to Appendix D.

Design and Materials

For the pilot test 50 general knowledge questions were selected from the German quiz web-page *http://wissen.de*. Questions on this web page have four short (one or two-word) multiple

exclusive answers. In the test, only three possible answers to each question were left, as one of the choices would usually be clearly incorrect. In choosing test questions I have tried to avoid the gender bias, which could result in inappropriate levels of under- or overconfidence for one gender, i.e. no questions that could be easier for men than women (e.g. sports, masculine hobbies) and vice versa were chosen. In the test students were asked to answer each of the 50 questions, and state their level of confidence in the correctness of their answer. Any number between 33% and 100% could be used to express subjects' confidence, where 33% meant that subjects did not know the correct answer, and were guessing, and 100% corresponded to being absolutely certain that the answer was correct.

In addition to measuring how well the subjects were calibrated, some personal data were collected: name, age, educational background, duration of studies, and nationality. In the final test students could also mark if they wanted to take part in the further experiments and, if answer was positive, submit their email. At the beginning of the pilot participants were informed that their personal data would be treated confidentially, and their identities would be used by the experimenter only for the purposes of determining the three winners. Thus, subjects' identity was revealed to other students only in the case of being one of the winners of the quiz, which was an honor to students. Test's instructions and design are based on the samples that were obtained from Dr. Briony Pulford (University of Leicester, School of Psychology) and Dr. Sabina Kleitman (University of Sidney, School of Psychology).

Based on the analysis of the pilot-test outcomes, a final test (*test-18*) was constructed from the 18 questions of the three difficulty levels: six hard, six medium difficulty, and six easy questions. Items were differentiated according to their difficulty on the basis of the number of correct answers to each of them from the whole group that participated in the pilot study. This methodology is described in the article of Pulford and Colman (1997), who suggest assigning questions to three difficulty categories, based on the total accuracy of the group in answering each question: 0-33% accuracy - hard questions, 34-66% - moderate difficulty, 67-100% easy questions. After the initial division, four questions have fallen in the category of hard questions (average accuracy 17.5%), 10 questions into category of medium difficulty questions (average accuracy 55.2%), and 36 of 50 questions turned to be easy (average accuracy 88.5%). As the category of hard-questions had not enough items, based on the idea that overconfidence is the most pronounced for hard questions (see Clarke, 1960; and Pitz, 1974), average overconfidence ratio over each of the medium difficulty questions was calculated and the two, having the highest overconfidence coefficient, were chosen to be included into hard-questions category. Thus six hard questions rather than four were obtained.

Characteristics of the final test in terms of the confidence, accuracy and the bias score are presented in the Table 1. Translation of *test-18* and instructions are included in Appendix H.

Table 1: Average confidence, accuracy and bias score for the three levels of question difficulty of the final overconfidence test (*test-18*)

	Hard		Medium		Easy	
	M	SD.	M	SD	M	SD
Confidence	67.90	6.64	65.01	9.01	97.43	2.12
Accuracy	26.00	16.00	62.33	2.34	100.00	0.00
Overconfidence	41.90	18.24	2.68	7.48	-2.57	2.12

5 RESULTS

Consistent with previous research, on average, subjects have proved to be overconfident: the bias score of the group on the *test-50* pointed at slight overconfidence (M = 4.47, SD = 7.34); recalculation of the bias score for the *test-18*[8] has increased the average overconfidence measure (M = 14.11, SD = 10.63). Appendix F presents data on the bias score and accuracy of all participants who took part in the pilot test for both *test-50* and *test-18*, and men and women separately. Average overconfidence of men for *test-50* is 3.33 (SD = 5.96), and for *test-18* it is 14.11 (SD = 10.70). Average overconfidence of women for *test-50* is 5.63 (SD = 8.47), and for *test-18* it is 14.12 (SD = 10.79). Noteworthy is the fact, that whereas for the complete *test-50* average overconfidence of men was slightly lower than that of women, after recalculating the overconfidence ratio for the questions chosen to comprise the final test (*test-18*), average bias score for both groups practically equalized. For the *test-50* correlation between accuracy and the bias score is found to be strong and significant, pointing at the decrease in overconfidence with the increase in accuracy (Pearson correlation (48) = -0.629, $p < 0.01$, one-sided); for the *test-18* this relationship is even stronger (Pearson correlation (48) = -0.823, $p < 0.01$, one-sided).

Overconfidence and experience

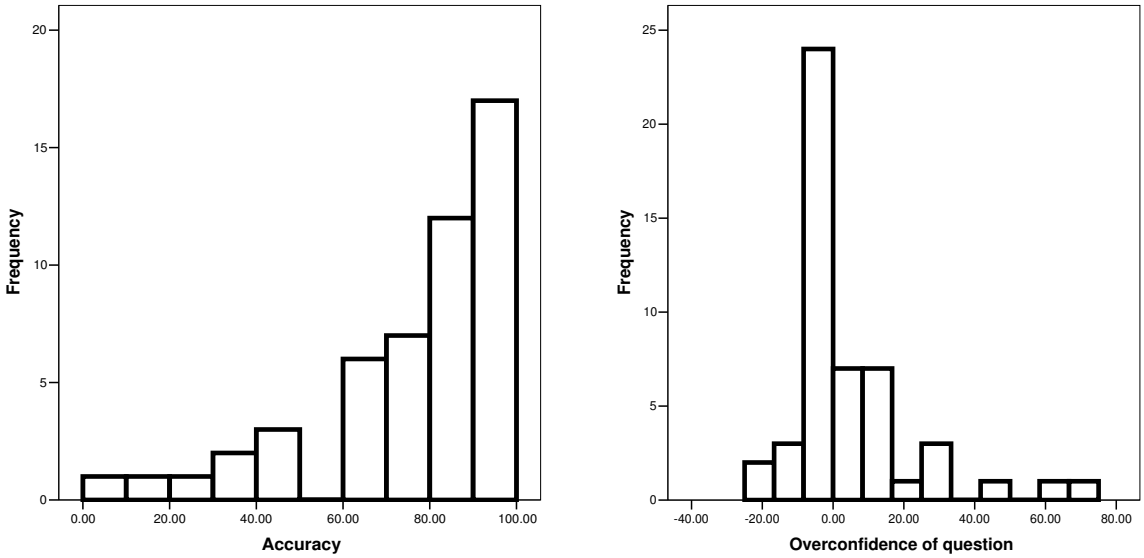
After obtaining the bias score for each individual participant of the pilot test, a check of the proposition that overconfidence of subjects changes with experience was conducted. The two

[8] Same subjects.

variables that were used as a proxy of subjects' experience are age and duration of study at the university, measured in semesters. From the graphical analysis of the scatterplots (Appendix B) no conclusions about linear relationship between the measures of experience and the bias score could be drawn for both *test-50* and *test-18*. Pearson's correlation analysis also has not detected any significant linear relationship between the variables of interest (see Appendix C). Based on these findings, I conclude that students of different age groups and being at different levels of progress with their studies can be recruited for the participation at the planned economic experiment.

Test-50 vs. Test-18: Accuracy and Confidence

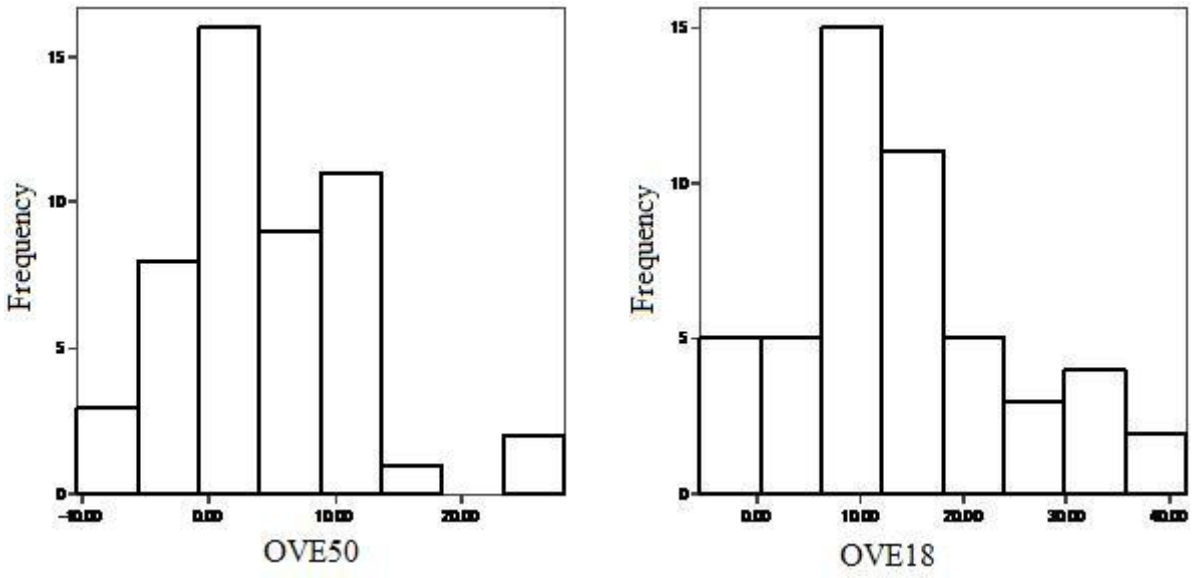
Analysis of the accuracy of the group for *test-50* revealed that even 72% of the questions have fallen in the category of easy questions (67-100% accuracy). See Figure 1(a). This test is distinguished by high precision, and inadequate to that precision confidence, consequently 58% of questions resulted in average underconfidence (see Figure 1(b). Appendix A (a) confirms, that the distribution of accuracy per question for the *test-50* has more mass on the right tail (skewness = -1.31), and the distribution of overconfidence per question is left-skewed (skewness = 1.86). This example illustrates the dangers of using the unbalanced to hard-easy effect test in economic research: by using *test-50* one can artificially create high levels of underconfidence in ones subjects[9].



a. b.

Figure 1: Distribution of accuracy (a) and overconfidence per question (b) in *test-50*

[9] Tests skewed in the direction of hard questions, can artificially create group overconfidence.



a. b.

Figure 2: Distribution of the bias score of participants per test: a. *test-50* and b. *test-18*

24% of subjects who completed *test-50* were found to be underconfident (see Figure 2(a); for *test-18* this number decreases to 8% (see Figure 2(b). Alongside with the decrease in the percentage of underconfident subjects, an increase in the range of the bias score of the participants is observed (from 38.60 to 47.23). There is also improvement in the symmetry of the distribution of the bias score (*test-50*: skewness = 0.73; *test-18*: skewness = 0.53). See Appendix A (b). The increase in bias score range is important for the future experiments as it leaves more room for finding subjects whose degree of overconfidence differs significantly.

5.1 Statistical Tests

In this section results of the statistical tests are presented that verify the success of categorization of the questions into three levels of difficulty for the *test-18*, and provide a sufficient basis to conclude that overconfidence is a robust phenomenon and not an artifact (Bar-Tal et al., 2001).

Confidence

I start by analyzing differences in the confidence levels of the subjects for the three difficulty levels of questions. On average subjects had the highest confidence for answering easy questions – 97.43%; the average confidence level for the hard questions was 67.90%, and for the medium questions – 65.01%. The performed Kruskal-Wallis H Test showed that the three levels of question difficulty resulted in significantly different from each other confidence levels (Chi-Square (2) = 11.617, $p < 0.01$), pointing out, that at least two of the three difficulty

levels were characterized by unequal confidence levels. Effect size is $\eta^2 = 0.856$, which means that 86% of the variance in the confidence assessments is due to the difficulty of the questions. Pairwise Mann-Whitney U tests revealed that there is no significant difference in the confidence for the medium and hard questions (U = 13.500, p = 0.470, two-sided); confidence in answering easy questions is significantly higher than in answering medium (U = 0.00, p < 0.01, one-sided) and hard questions (U = 0.00, p < 0.01, one-sided).

Accuracy

Average accuracy levels for the *test-18* were: 26% for hard questions, 62.33% for medium, and 100% for easy questions. Kruskal-Wallis H Test indicates that the difficulty level of questions significantly affected accuracy of the answers (Chi-Square (2) = 15.760, p = 0.00); effect size is $\eta^2 = 0.926$. A series of Mann Whitney U tests were carried out. These tests show that there is a significant difference in accuracy for answering three categories of questions: medium questions tend to outperform in accuracy hard questions (U = 0.00, p < 0.01, one-sided); accuracy for answering easy questions significantly exceeds the accuracy of medium (U = 0.00, p < 0.01, one-sided), and hard questions (U = 0.00, p < 0.01, one-sided). These results prove that the division of questions into three difficulty levels is successful.

Overconfidence

Table 1 demonstrates that participants exhibit overconfidence for two levels of question difficulty (hard questions: BS = 41.90; medium questions: BS = 2.68) and underconfidence for the third one (easy questions: BS = -2.57). This is in line with the previous research that found hard questions to be the most prone to overconfidence, and easy questions to be often subject to underconfidence. The bias scores for easy and hard questions differ significantly from zero (easy questions: Wilcoxon signed rank test T = 2.097, p < 0.05, two-sided; hard questions: Wilcoxon T = 2.097, p <0.05, two-sided). However, for the medium difficulty questions the null hypothesis of the equality of the bias score to zero cannot be rejected (Wilcoxon T = 0.419, p = 0.675, two-sided). It can be concluded that the medium difficulty questions produced on average the bias score which was the most indistinguishable from the perfect calibration score of zero. To examine the existence of the hard-easy effect I first test a joint hypothesis of the equality of the levels of overconfidence generated by three levels of questions' difficulty versus the alternative, that some difficulty levels produced more overconfidence than the others. The null hypothesis is rejected at a high level of significance (Chi-Square (2) = 12.117, p < 0.01). Effect size is $\eta^2 = 0.783$. Mann-Whitney U test, performed on each pair of the three levels of the bias score, confirmed the existence of the hard-easy effect. Subjects showed significantly higher overconfidence for the hard questions than for the medium (Mann-Whitney U = 0.00, p < 0.01,

one-sided) and easy ones (Mann-Whitney U = 0.00, $p < 0.01$, one-sided); overconfidence for the medium questions was slightly higher than for the easy questions (Mann-Whitney U = 9.50, $p < 0.1$, one-sided).

Gender Differences

Test-50:

Males were slightly less overconfident than females for the *test-50* (men BS: M = 3.33, SD = 5.96; women BS: M = 5.63, SD = 8.47), however this difference was not significant (t(48) = -1.109, p = 0.27, two-sided; effect size $\eta^2 = 0.025$) (see Appendix F). Male subjects achieved higher accuracy for the *test-50* than female subjects (men: M = 78.80, SD = 5.45; women: M = 73.52, SD = 6.72), and this difference is significant (t(48) = 3.053, $p < 0.01$, two-sided). Effect size is $\eta^2 = 0.163$, which points out that 16.3% of the variance in accuracy was gender dependent. Male subjects have also shown higher confidence in answering questions of *test-50*, than female subjects (men: M =82.13, SD = 4.93; women: M = 79.07, SD = 6.70), this difference is found to be significant (t(48) = 1.840, $p < 0.05$, one-sided); effect size $\eta^2 = 0.069$. The fact that about 16% of variation in accuracy and 7% in confidence is gender dependent is not satisfactory because there is more gender bias in the overconfidence test than it was expected. Correlation between overconfidence and accuracy is strong and significant for both genders (men: Pearson's Correlation (23) = -0.630, $p < 0.01$, one-sided; women: Pearson's Correlation (23) = -0.625, $p < 0.01$, one-sided).

Test-18:

Both genders have shown almost equal overconfidence for the *test-18* (male BS: M = 14.11, SD = 10.70; female BS: M = 14.12, SD = 10.79; t(48) = -0.002, p = 0.998, two-sided; effect size $\eta^2 = 0.00$) (see Appendix F). Overconfidence for the *test-50* was significantly lower than for the *test-18*. Male subjects have slightly higher accomplishments in terms of accuracy than female subjects (men: M = 63.78, SD = 9.64; women: M = 61.78, SD = 10.43), although this difference is found to be insignificant (t(48) = 0.704, p = 0.485, two-sided); effect size $\eta^2 = 0.010$. Male subjects were slightly more confident in answering questions of *test-18* (men: M = 77.40, SD = 5.21; women: M = 74.87, SD = 5.20), however this difference is insignificant (t(48) = 1.37, p = 0.176, two-sided); effect size $\eta^2 = 0.037$. Compared to test-50, test-18 has very low amount of variation in confidence and accuracy that is gender dependent. Test of the difference in overconfidence between men and women for the three levels of question difficulty has shown that both groups have expressed similar biases in answering the test and that the encountered differences were not significant (hard questions: t(48) = 0.085, p = 0.933, two-sided; medium questions: t(48) = 0.354, p = 0.725, two-sided; easy questions: t(48) =

0.737, p = 0.465, two-sided). Correlation between overconfidence and accuracy is strong and significant for both genders (men: Pearson's Correlation (23) = -0.847, p < 0.01, one-sided; women: Pearson's Correlation (23) = -0.810, p < 0.01, one-sided).

6 SECOND EXPERIMENT

To check if the results obtained by using *test-18* were replicable, namely the average group degree of overconfidence, the obtained categorization into three difficultly levels and controlling for gender bias, the experiment was repeated with the students of the target group: those enrolled into different disciplines of social sciences. In this subsection I will also estimate the reliability of my scale.

A second experiment was conducted on the 14th June, 2008 at Christian-Albrechts University of Kiel. Subjects were given approximately 15 minutes time to fill in the final, 18 questions, overconfidence test (*test-18*) at the end of the lecture on Economics of Risk and Uncertainty. As in the pilot test, three monetary prizes were offered for the participants who got the most questions right. A total of 37 tests were completed, of them 3 had no personal information and were not included in the further analysis. Participants of the test aged from 22 to 31 years (M = 26.06, SD = 2.62), and have studied on average 9.10 semesters (SD = 2.60). Of the 34 participants 21 were males (age: M = 25.95, SD = 2.64), and 13 were females (age: M = 26.23, SD = 2.68). The majority of the subjects were Germans (86%). All participants were students of social sciences, of them 26 studied economics, seven studied management, and one other social sciences. For information about subjects' age and duration of studies refer to Appendix E. Consistent with previous research, on average, subjects were prone to overconfidence (M = 10.41, SD = 9.26). Average group overconfidence on test-18 obtained from the experiment on the 6th, June and on the 19th, May did not significantly differ from each other (t(82) = 1.649, p = 0.103, two-sided; size effect $\eta^2 = 0.032$). Men on average were slightly more overconfident (M = 10.68, SD = 9.81) than women (M = 9.98, SD = 8.68), however this difference was found to be insignificant. Appendix G presents data on the bias score of all participants who took part in the pilot, and men and women separately. Just as in the pilot test, correlation coefficient between age and overconfidence (Pearson coefficient (32) = 0.189, p = 0.142, one-sided), and semester and overconfidence (Pearson coefficient (32) = -0.054, p = 0.388, one-sided) is small and insignificant. Correlation between the accuracy and the bias score is strong and significant, pointing at the decrease in overconfidence with the increase in accuracy (Pearson correlation (332) = -0.731, p < 0.01). After division of questions into three difficulty levels characteristics of each category, in terms of confidence, accuracy and the bias score, were calculated (see Table 2).

Table 2: Characteristics of the three levels of question difficulty of the *test-18* from the experiment on 14.06.09

	Hard		Medium		Easy	
	M	SD	M	SD	M	SD
Accuracy	22.35	12.06	52.94	9.49	95.38	7.76
Confidence	55.75	7.67	53.95	11.67	93.52	8.74
BS	33.40	19.24	1.01	11.38	-1.86	2.42

Accuracy

Subjects' average accuracy level for answering hard questions was 22.35%, 52.94% for medium and 95.38% for easy questions. Kruskal-Wallis H Test shows that the difficulty level of questions had significant impact on the accuracy of answers (Chi-Square (2) = 15.065, $p < 0.01$); effect size $\eta^2 = 0.920$. Pairwise comparisons, performed using the Mann-Whitney U test, revealed that the accuracy for answering the medium difficulty questions significantly exceeds the accuracy for answering hard questions ($U = 0.50$, $p < 0.01$, one-sided); accuracy for answering easy questions significantly exceeds accuracy for medium questions ($U = 0.00$, $p < 0.01$, one-sided) and hard questions ($U = 0.00$, $p < 0.01$, one-sided).

Confidence

On average subjects had the most confidence for answering easy questions 93.52% (SD = 8.74); confidence levels for the hard and medium difficulty questions were correspondingly 55.75% (SD = 7.67) and 53.95% (SD = 11.67). The Kruskal-Wallis H Test demonstrated that the three difficulty levels differed significantly from each other with regard to confidence (Chi-Square (2) = 12.158, $p < 0.01$) (see Appendix G). Effect size is $\eta^2 = 0.824$, which means that 82% of the variance in confidence assessments is due to the difficulty of questions. Pairwise comparison revealed that the confidence for easy questions was significantly higher than the confidence for medium and hard questions (both: Mann-Whitney $U = 0.00$, $p < 0.01$, one-sided); however there was no significant difference in the confidence levels for the hard and medium questions (Mann-Whitney $U = 15.00$, $p = 1.00$, two-sided).

Overconfidence

The overconfidence level for the hard questions was the highest (M = 33.40, SD = 19.24), the medium difficulty questions produced almost no overconfidence (M = 1.01, SD = 11.38),

whereas easy questions resulted on average in underconfidence (M = -1.86, SD = 2.42). The three difficulty levels of questions differed significantly in terms of the produced bias score (Chi-Square (2) = 9.079, $p < 0.01$), size effect $\eta^2 = 0.659$. Mann-Whitney U analysis showed subjects showed significantly higher overconfidence for the hard questions than for the medium (Mann-Whitney U = 2.00, $p < 0.01$) and easy questions (Mann-Whitney U = 0.00, $p < 0.01$), which is in line with the previous research. Overconfidence levels for medium and easy questions, on average, were not significantly different from each other (Mann-Whitney U = 18.00, p = 0.334, one-sided). The bias score for the hard questions was significantly higher than zero (Wilcoxon Signed Rank T = 2.097, $p < 0.05$, one-sided); for the easy and medium difficulty questions the null hypothesis of the equality of the bias score to zero cannot be rejected (easy questions: Wilcoxon Signed Rank T = 1.606, p = 0.108, two-sided; medium questions: Wilcoxon Signed Rank T = 0.00, p = 1.00, two-sided). For this group of the participants, easy and medium difficulty questions produced on average the bias score which was the most indistinguishable from the perfect calibration score of zero.

Gender differences

No significant difference between male and female participants in terms of overconfidence was found (t(32) = 0.211, p = 0.834, two-sided; effect size $\eta^2 = 0.001$) (see Appendix G). Although men, on average, were less accurate than women this difference is not significant (t(32) = -0.524, p = 0.604, two-sided; effect size $\eta^2 = 0.009$). The difference in average confidence across all items of the test between male and female participants is insignificant (t(32) = -0.53, p = 0.600, two-sided; effect size is $\eta^2 = 0.009$). No significant difference in overconfidence is found between male and female subjects for the three levels of question difficulty (hard questions: t(32) = 0.042, p = 0.967, two-sided; medium questions: t(32) = -0.357, p = 0.723, two-sided; easy questions: t(32) = 1.468, p = 0.152, two-sided). Correlation between overconfidence and accuracy is strong and significant for both genders (men: Pearson's Correlation (19) = -0.653, $p < 0.01$, one-sided; women: Pearson's Correlation (11) = -0.883, $p < 0.01$, one-sided).

Reliability

According to DeCoster (2000), a scale can be called reliable (possess internal consistency) "if repeated measurements under the same circumstances tend to produce the same results". A common way to estimate reliability of an instrument is to calculate Cronbach's alpha. Moss et al (1993) state, that a generally acceptable value of coefficient alpha equals 0.6; however the more recognized threshold is 0.7. These values of alpha are considered to be optimal for the use in social research. For my instrument three values of alpha were estimated: alpha for the

test confidence equaled 0.79, alpha for the test accuracy – 0.54, and alpha for the bias score – 0.68. Values of the calculated alphas were either close or exceeded the threshold values. A somewhat lower degree of alpha for the accuracy dimension resulted from low variance in answering easy questions. Easy questions cannot be removed from the test, in the desire to improve its reliability, as a good instrument should not only have a reasonable internal consistency (reliability) but also a "meaningful content coverage" (Schmitt, 1996). Based, on the calculated values of Cronbach's alpha, it can be concluded that the developed instrument possesses good internal consistency (reliability).

7 CONCLUSIONS

In this paper results of the two experiments, aimed at the development of the instrument (test) that would enable the construction of the comprehensive measure of individual overconfidence, are presented. Desired instrument, which is to be used in economic overconfidence experiments, should allow assessment of the differences between the subjects with respect to their degree of overconfidence and minimize the measurement error.

After carrying out the analysis of some of the instruments used in foregoing experimental research, there were good reasons to suspect that overconfidence was previously measured inadequately. The principal steps needed to improve the instrument (test) were: 1) choice of another test-format (multiple choice discrete propositions' tasks instead of confidence intervals estimation), 2) balancing the test for the hard-easy effect, and 3) controlling for gender and country bias. Instrument was obtained in a two-stage procedure in which a pilot test was used to assess questions' difficulty, based on the groups' accuracy in answering each of the initial 50 items. Then six questions of the three difficulty types (hard, medium, and easy) were selected for the final test. The second experimental phase was aimed at verification of replicability of results, namely of the average degree of group overconfidence, the obtained categorization into three difficultly levels and of controlling for the gender bias. Both experiments were conducted with the students enrolled into different disciplines of social sciences. The two experimental sessions were administered and subjects were offered a reward, on the basis of competition in test accuracy. As in previous experimental work, subjects on average proved to be overconfident.

Evidence was found for the significant effect of the question difficulty on the overconfidence measure and for the existence of the gender bias. Hard questions produced significantly higher levels of overconfidence than medium-difficulty and easy questions, which in turn resulted in underconfidence. Analysis of the groups' accuracy on answering initial test (*test-*

50) revealed that even 72 percent of the questions fell in the category of easy questions. Thus, by using initial *test-50* to measure subjects' overconfidence, one would artificially create high levels of underconfidence in ones subjects. Statistical analysis confirmed that in both experimental sessions the three types of questions, that comprised the final test, significantly differed from each other in terms of the produced confidence, accuracy and overconfidence. This result verified the success of categorization of questions into three levels of difficulty in the overconfidence measurement instrument. Average group overconfidence measures on *test-18*, obtained from both experimental sessions, did not differ significantly from each other. Instrument's internal consistency (reliability), assessed as the value of the Cronbach's alpha, was found to be good and acceptable for the use in social research.

Combining all levels of questions' difficulty, both genders expressed overconfidence that did not differ significantly from each other. It can be concluded, that for the created instrument (*test-18*), gender is not associated with overconfidence: first, there were no significant differences between male and female subjects' bias scores and, second, no significant difference in overconfidence was found between male and female subjects for the three levels of question difficulty. There was also almost no variance in confidence and accuracy that was gender dependent. By contrast, for the initial instrument (*test-50*) as much as 16 percent of variance in accuracy and 7 percent of variance in confidence was explained by gender.

Based on the analysis of the data obtained from both phases of the instrument construction, and in the light of the importance of employment of a reliable measure to assess subjects' overconfidence for the validity of the results of economic experiments, it can be concluded that a better instrument was developed for the use in planned experiments, suitable for evaluation of individual differences in terms of the degree of overconfidence.

REFERENCES

Adams, J. K. (1957). A confidence scale defined in terms of expected percentages. The American Journal of Psychology, Vol. 70(3), p. 432-436.

Adams, P. A., Adams, J. K., (1958), Training in confidence-judgments. The American Journal of Psychology, Vol. 71(4), p. 747-751.

Allwood, C.M., Granhag, P. A., Jonsson, A. C., (2006), Child witnesses' metamemory realism. Scandinavian Journal of Psychology, Vol. 47(6), p 461-447.

Alpert, M., Raiffa, H., (1982), A progress report on the training of probability assessors. In Kahneman, D., Slovic, P., Tversky, A. (Eds.), Judgment under Uncertainty: Heuristics and Biases, Cambridge University Press, p. 294-305.

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

Barber, B., Odean, T., (2001), Boys will be boys: Gender, overconfidence and common stocks investments. Quarterly Journal of Economics, Vol. 116, p. 261-292.

Barber, B. M., Odean, T., (2002), Online investors: do the slow die first? Review of Financial Studies, Vol. 15(2), p. 455-487.

Bar-Hillel, M., (2001), Subjective probability judgments. In Smelser, N.J. and Baltes, D.B. (Eds.), International Encyclopedia of the Social & Behavioral Sciences, Amsterdam: Elsevier Science Ltd., p. 15247-15251.

Bar-Tal Y., Sarid A., Kishon-Rabin L., (2001), A test of overconfidence phenomenon using audio signals. The Journal of General Psychology, Vol. 128(1), p. 76-80.

Baron, R. A., (2000), Psychological Perspectives on Entrepreneurship: Cognitive and Social Factors in Entrepreneurs' Success. Current Directions in Psychological Science, Vol. 9(1), p. 15-18.

Benos, A., (1998), Aggressiveness and Survival of Overconfident Traders. Journal of Financial Markets, Vol. 1, p. 353-383.

Bernard V. L., Thomas J., (1989), Post-earnings-announcement drift: delayed price response or risk premium. Journal of Accounting Research, Vol. 27, p. 1–36.

Bernard, V. L., Thomas J., (1990), Evidence that stock prices do not fully reflect the implications of current earnings for future earnings. Journal of Accounting and Economics, Vol. 13(4), p. 305–340.

Biais, B., Hilton, D., Mazurier, K., Pouget, S., (2005), Judgmental overconfidence, self-monitoring and trading performance in an experimental financial market. Review of economic studies, Vol. 72(2), p. 287-312.

Blavatsky, P., (2009), Betting on own knowledge: Experimental test of overconfidence. Journal of Risk and Uncertainty, Vol. 38, p. 39-49.

Bohner, G., Rank, S., Reinhard, M. A., Einwiller, S., Erb, H. P., (1998), Motivational determinants of systematic processing: expectancy moderates effects of desired confidence on processing effort. European Journal of Social Psychology, Vol. 28, p. 185-206

Caballé, J., Sákovics J., (2003), Speculating against an overconfident market. Journal of Financial Markets, Vol. 6, p. 199-225.

Cambridge, R. M., Shreckengost, R. C., (1978), Are you sure? The subjective probability assessment test. Langley, VA: Office of Training, Central Intelligence Agency.

Camerer, C., (1995), Individual decision making. In Kagel, J. H., Roth, A. (Eds.), Handbook of experimental economics, Princeton University press, p. 587-703.

Clarke, F. R., (1960), Confidence Ratings, Second-Choice Responses, and Confusion Matrices in Intelligibility Tests. Journal of the Acoustical Society of America, Vol. 32(1), p. 35-46.

Conger, R. F., Wolstein, Ch. R., (2004), Managing overconfidence in pricing. Emphasis, Vol. 2, p. 10-13.

Daniel, K., Hirshleifer D., Subrahmanyam A., (1998), Investor psychology and security market under-and overreactions. Journal of Finance, Vol. 53(6), p 1839-1885.

Daniel, K. D, Hirshleifer, D., Subrahmanyam, A., (2001), Overconfidence, arbitrage, and equilibrium asset pricing. Journal of Finance, Vol. 56 (3), p. 921–965.

Deaves, R., Lüders, E., Luo, G. Y., (2009), An Experimental Test of the Impact of Overconfidence and Gender on Trading activity. Review of finance, Vol. 13(3), p. 555–575.

DeCoster, J., (2005), Scale Construction Notes. Retrieved on 04.30.2010 from http://www.stat-help.com/notes.html

De Bondt, W. F. M., Thaler R., (1984), Does the Stock Market Overreact? Journal of Finance, Vol. 40(3), p. 793-805.

De Long, J. B., Shleifer, A., Summers, L. H., Waldmann, R. J., (1991), The Survival of Noise Traders in Financial Markets. The Journal of Business, Vol. 64(1), p. 1-19.

Fenton-O'Creevy, M., Nicholson, N., Soane, E., Willman, P., (2003), Trading on Illusions: Unrealistic Perceptions of Control and Trading Performance. Journal of Occupational and Organizational Psychology, Vol. 76, p. 53–68.

Fischhoff, B., Slovic, P., Lichtenstein, S., (1977), Knowing with certainty: The appropriateness of extreme confidence. Journal of Experimental Psychology: Human Perception and Performance, Vol. 3, p. 552-564.

Fischhoff, B., (1982), Debiasing. In Kahneman, D., Slovic, P., Tversky, A. (Eds.), Judgment under Uncertainty: Heuristics and Biases, Cambridge University Press, p. 422-444.

Gervais, S., Odean, T., (2001), Learning to be Overconfident. Review of Financial Studies, Oxford University Press for Society for Financial Studies, Vol. 14(1), p. 1-27.

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

Glaser, M., Langer, T., Weber, M., (2005), Overconfidence of Professionals and Lay Men: Individual Differences within and between Tasks? Working Paper, Mannheim.

Glaser, M., Weber, M., (2007), Overconfidence and trading volume. The Geneva Risk and Insurance Review, Vol. 32(1), p. 1-36.

Hazard, T. H., Peterson, C. R., (1973), Odds versus probabilities for categorical events. Tech. Rep. No. 73-2, McLean, VA: Decisions and Designs, Inc.

Hirshleifer D., Subrahmanyam, A., Titman S., (1994), Security analysis and trading patterns when some investors receive information before others. Journal of Finance, Vol. 49(5), p. 1665-1698.

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

Hynes, M., Vanmarcke, E., (1976), Reliability of embankment performance predictions. Proceedings of the ASCE Engineering Mechanics Division Specialty Conference, Waterloo, Ontario, Canada: University of Waterloo Press.

Johnson, D. D. P., McDermott, R., Barrett E. S., Cowden, J., Wrangham R., McIntyre, M.H., Rosen, S. P., (2006), Overconfidence in wargames: experimental evidence on expectations, aggression, gender and testosterone. Proceedings: Biological Sciences, Vol. 273(1600), p. 2513-2520.

Kahneman D., Riepe, M. W., (1998), Aspects of Investor Psychology. Journal of Portfolio Management, Vol. 24(4), p. 52-65.

Keasey, K., Watson, R., (1989), Consensus and accuracy in accounting studies of decision making: A note on a new measure of consensus. Accounting, Organizations and Society, Vol. 14, p. 337-345.

Kim, A. K., Nofsinger, J. R., (2003), The behavior and performance of individual investors in Japan. Working paper.

Kirchler, E., Maciejovsky, B., (2002), Simultaneous over- and underconfidence: evidence from experimental asset markets. Journal of Risk and Uncertainty, Springer, Vol. 25(1), p. 65-85.

Klayman, J., Soll., J. B., Gonzáles-Vallejo, C., Barlas, S., (1999), Overconfidence: it depends on how, what, and whom you ask. Organizational Behavior and Human Decision Processes, Vol. 79(3), p. 216-247.

Koriat A, Lichtenstein S, Fischhoff B., (1980), Reasons for confidence. Journal of Experimental Psychology: Human Learning and Memory, Vol. 6, p. 107-18.

Kyle, A., Wang, F. A., (1997), Speculation duopoly with agreement to disagree: can overconfidence survive the market test? Journal of Finance, Vol. 52(5), p. 2073-2090.

Lakonishok, J., Shleifer, A., Vishny, R. W., (1992), The impact of institutional trading on stock prices. Journal of Financial Economics, Vol. 32(1), p. 23-43.

Langer, E. J., (1982), The illusion of control. In: Kahneman, D., Slovic, P., Tversky A. (Eds.), Judgment under uncertainty: heuristics and biases, Cambridge University Press, p. 231-338.

Loughran, T., Ritter, J. R., (1995), The new issues puzzle. Journal of Finance, Vol. 50(1), p. 23-51.

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

Lichtenstein, S., Fischhoff, B., (1981), The effects of gender and instructions on calibration (Decision Research Report 81-5). Eugene OR: Decision Research.

Lichtenstein, S., Fischhoff, B., and Phillips, L. D., (1977), Calibration of probabilities: The state of the art. In Jungermann, H., deZeeuw, G. (Eds.), Decision making and change in human affairs, Amsterdam: D. Reidel, p. 275-324.

Lichtenstein, S., Fischhoff, B., Phillips, L. D., (1982), Calibration of probabilities: the state of the art to 1980. Judgment under Uncertainty: Heuristics and Biases. In Kahneman, D., Slovic, P., Tversky, A. (Eds.), Judgment under Uncertainty: Heuristics and Biases, Cambridge University Press, p. 306-334.

Menkhoff, L., Schmidt, U., Brozynski, T., (2006), The impact of experience on risk taking, overconfidence, and herding of fund managers: Complementary survey evidence. European Economic Review, Vol. 50(7), p. 1753-1766

Menkhoff, L., Schmeling, M., Schmidt, U., (2006), Does Professionalism Consistently Affect Portfolio Biases? Working Paper.

Michaely, R., Thaler, R. H., Womack, K. L., (1995), Price Reactions to Dividend Initiations and Omissions: Overreaction or Drift? Journal of Finance, Vol. 50(2), p. 573-608.

Moore, P. G., (1977), The manager's struggle with uncertainty. Journal of the Royal Statistical Society, Vol. 140, p. 129-165.

Moss, S., Patel, P., Prosser, H., Goldber, D., Simpson, N., Rowe, S., Lucchino, R., (1993), Psychiatric morbidity in older people with moderate and severe learning disability. I: Development and reliability of the patient interview (PAS–ADD). British Journal of Psychiatry, Vol. 163, p. 471-480.

Neale, M. A., Bazerman, M. H., (1990), Cognition and rationality in negotiation. New York: The Free Press, p. 240.

Nickerson, R. S., McGoldrick, C. C., (1965), Confidence ratings and level of performance on a judgmental task. Perceptual & Motor Skills, Vol. 20, p. 311-316.

Nöth, M., Weber, M., (2003), Information Aggregation with Random Ordering: Cascades and Overconfidence. The Economic Journal, Vol. 113, p. 166-189.

Oberlechner, T., Osler, C.L., (2003), Overconfidence in currency markets. Working Paper.

Odean, T., (1998), Volume, volatility, price and profit when all traders are above average. Journal of Finance, Vol. 53(6), p. 1887 -1934.

Odean, T., (1999), Do investors trade too much? American Economic Review, Vol. 89(5), p. 1278-1298.

Oechssler, J., Schmidt, C., Schnedler, W., (2007), Asset bubbles without dividends - an experiment. Working Paper Nr. 07-01, University of Mannheim.

Oskamp, S., (1962), The relationship of clinical experience and training methods to several criteria of clinical prediction. Psychological Monographs, Vol. 76(547), p. 1-28.

Pitz, G.F., (1974), Subjective probability distributions for imperfectly known quantities. In Gregg, L. W. (Ed.), Knowledge and Cognition. New York: Wiley, p. 29-41.

Powel, W. D., Bolich, C., (1993), Changing predictions based on knowledge of past performance in preschool and young grades-school. Child Study Journal, Vol. 23(3), p. 209.

Pulford, D. B., (1996), Overconfidence in Human Judgment, PhD thesis. Department of Psychology, University of Leicester.

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

Ronis, D. L., Yates, J. F., (1987), Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. Organizational Behavior and Human Decision Processes, Vol. 40, p. 193-218.

Russo, J. E., Schoemaker, P. J., (1992), Managing overconfidence. Sloan Management Review, Vol. 33, p. 7-17.

Scheinkman, J. A., Xiong, W., (2003), Overconfidence and speculative bubbles. Journal of Political Economy, Vol. 111, p. 1183-1219.

Shiller, R. J., (2000), Irrational Exuberance. Princeton, N.J.: Princeton Univ. Press, 344 p.

Sieber, J. E., (1979), Confidence estimates on the correctness of constructed and multiple-choice responses. Contemporary Educational Psychology, Vol. 4, p. 272-287.

Sniezek, J. A., Paese, P. W., Switzer, F. S., (1990), The effect of choosing on confidence in choice. Organizational Behavior and Human Decision Processes, Vol. 46, p. 264-282.

Staël von Holstein, C. A. S., (1972), Probabilistic forecasting: An experiment related to the stock market. Organizational Behavior and Human Performance, Vol. 8, p. 139-158.

Statman, M., Thorley, S., Vorkink, K., (2006), Investor overconfidence and trading volume. Review of Financial Studies, Vol. 19, p. 1531–1565.

Stotz, O., von Nitzsch, R., (2005), The perception of control and the levels of overconfidence: evidence from analyst earnings estimates and prices targets. The Journal of Behavioral Finance, Vol. 6(3), p. 121-128.

Sümer, N., Özkan, T., Lajunen, T., (2006), Asymmetric relationship between driving and safety skills. Accident Analysis and Prevention, Vol. 38(4), p. 703-711.

Svenson, O., (1981), Are we all less risky and more skilful than our fellow drivers? Acta Psychologica, Vol. 47(2), p. 143-148.

Taylor, S. E., and Brown, J. D., (1988), Illusion and well-being: A social psychological perspective on mental health. Psychological Bulletin, Vol. 103, p. 193-210.

Tversky, A., Kahneman, D, (1982), Judgment under Uncertainty: Heuristics and Biases. In Kahneman, D., Slovic, P., Tversky, A. (Eds.), Judgment under Uncertainty: Heuristics and Biases, Cambridge University Press, p. 3-20.

Wagenaar, W. A., Keren, G., (1986), Does the expert know? The reliability of predictions and confidence ratings of experts. In E. Hollnagel, G. Maneini, and D. D. Woods (Eds.), Intelligent decision support in process environments, Berlin: Springer, p. 87-107.

Weinstein, N. D., (1980), Unrealistic optimism about future life events. Journal of Personality and Social Psychology, Vol. 39, p. 806-820.

Zakay, D., Glicksohn, J., (1992), Overconfidence in a multiple-choice test and its relationship to achievement. Psychological Record, Vol. 42(4), p 519-524.

APPENDIX A:

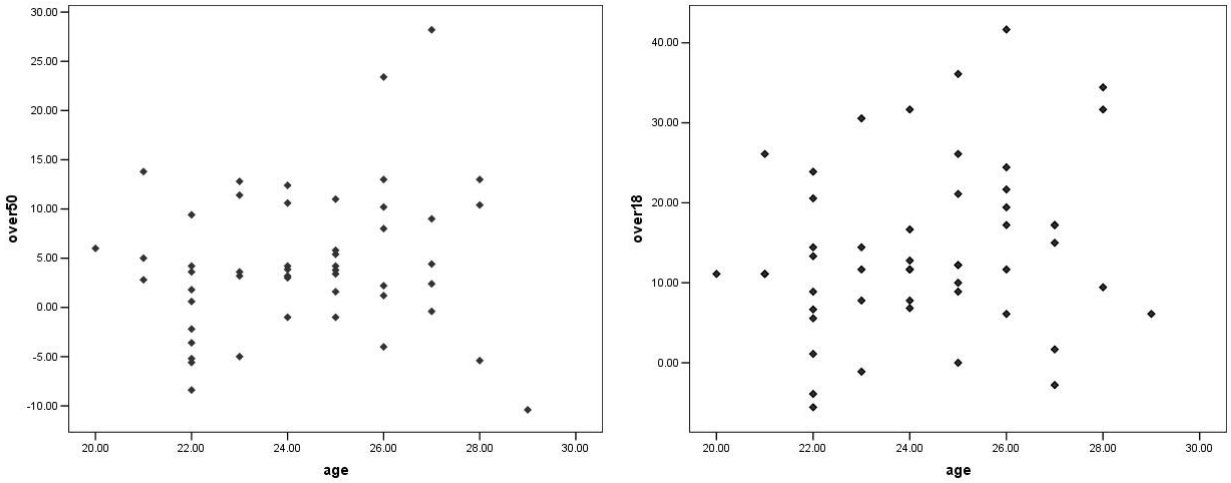
a: Skewness of the accuracy parameters and overconfidence scores per question for the *test-50*

		Accuracy	BS
N		50	50
Skewness		-1.310	1.855
Std. err. of skewness		0.337	0.337
Range		92.00	91.40
Min		8.00	-22.80
Max		100.00	68.60
Percentiles:	25	63.50	-4.60
	50	84.00	-1.10
	75	94.50	11.60

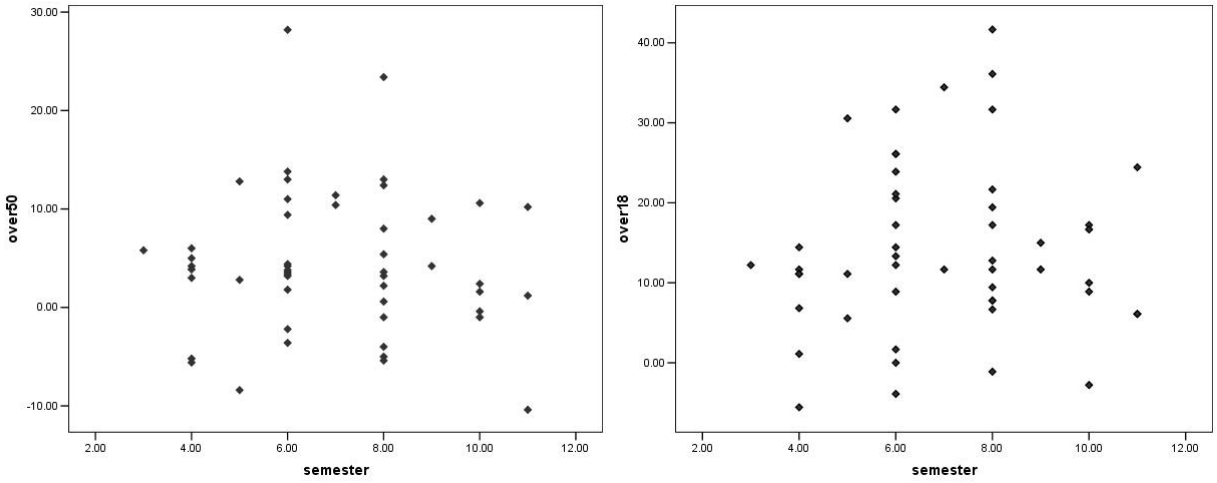
b: Comparison of *test-50* and *test-18* in terms of overconfidence

		BS50	BS18
N		50	50
Mean		4.47	14.11
SD		7.34	10.63
Skewness		0.726	0.525
Std. err. of skewness		0.337	0.337
Range		38.60	47.23
Min		-10.40	-5.56
Max		28.20	41.67
Percentiles:	25	0.35	7.54
	50	3.70	11.95
	75	9.60	20.70

APPENDIX B: SCATTERGRAMS OF THE EXPERIENCE MEASURES RELATIONSHIP TO OVERCONFIDENCE (a. *test-50* and *test-18* age vs. bias score, and b. *test-50* and *test-18* semester vs. bias score).



a.



b.

APPENDIX C: PEARSON'S TEST (DF. = 48) FOR CORRELATION RESULTS

		Semester	Age
OVE50	Correlation Coefficient	-0.045	0.148
	Sig. (one-sided)	0.377	0.152
OVE18	Correlation Coefficient	0.078	0.194
	Sig. (one-sided)	0.312	0.088

APPENDIX D: AGE AND STUDY DURATION INFORMATION OF THE PILOT TEST ON 19.05.2008

	N	M	SD	Min	Max
Age	50	24.32	2.20	20	29
Male age	25	24.48	2.43	20	29
Female age	25	24.16	1.97	21	27
Semester	50	6.98	2.11	3	11
Male semester	25	7.00	2.27	3	11
Female semester	25	6.96	1.99	4	11

APPENDIX E: AGE AND STUDY DURATION INFORMATION OF THE PILOT TEST ON 14.06.2008

	N	M	SD	Min	Max
Age	34	26.06	2.62	22	31
Male age	21	25.95	2.64	22	31
Female age	13	26.23	2.68	22	30
Semester	30	9.10	2.60	4	15
Male semester	20	9.10	2.97	4	15
Female semester	10	9.10	1.79	6	11

APPENDIX F: INFORMATION ON OVERCONFIDENCE AND ACCURACY OF THE PARTICIPANTS OF PILOT ON 19.05.2008

Overconfidence

Pilot Test 50					
OBS	Group	Mean	SD	Mini	Max
50	All	4.48	7.34	-10.40	28.20
25	Female	5.63	8.47	-8.40	28.20
25	Male	3.33	5.96	-10.40	13.00
	Male vs. female diff.	-2.298 (0.273)			
Pilot Test 18					
OBS	Group	Mean	SD	Min	Max
50	All	14.11	10.63	-5.56	41.67
25	Female	14.12	10.79	-5.56	41.67
25	Male	14.11	10.70	-3.89	36.11
	Male vs. female diff.	-0.007 (0.998)			

APPENDIX F - CONTINUATION:

Accuracy

Pilot Test 50					
OBS	Group	Mean	SD	Min	Max
50	All	76.16	6.61	58	90
25	Female	73.52	6.72	58	84
25	Male	78.80	5.45	66	90
	Male vs. female diff.	5.28 (0.004)			
Pilot Test 18					
OBS	Group	Mean	SD	Min	Max
50	All	62.78	9.99	38.89	83.33
25	Female	61.78	10.43	38.89	77.78
25	Male	63.78	9.64	44.44	83.33
	Male vs. female diff.	2.00 (0.485)			

APPENDIX G: INFORMATION ON OVERCONFIDENCE AND ACCURACY OF THE PARTICIPANTS OF PILOT ON 14.06.2008

Overconfidence

OBS	Group	M	SD	Min	Max
34	All	10.41	9.26	-6.28	30.00
13	Female	9.98	8.68	-3.44	28.94
21	Male	10.68	9.81	-6.28	30.00
	Male vs. female diff.	0.700 (0.604)			

Accuracy

OBS	Group	M	SD	Min	Max
34	All	60.46	9.35	38.89	77.78
13	Female	61.54	9.48	38.89	77.78
21	Male	59.79	9.45	38.89	77.78
	Male vs. female diff.	-1.750 (0.834)			

General Knowledge Questionnaire

Below you will be presented with some general knowledge questions. Imagine that you are taking part in a game, like "Trivial Pursuit" or "Who wants to be a Millionaire?", and you have to choose the correct answer from the three given alternatives. A person who answers the most questions right will get a 30 EUR prize. The second place will be awarded by the 20 EUR prize, and the third place by 10 EUR. You will be paid next week!

1) Please circle ONLY ONE of three given answers. Only one of them is correct.

2) When you have made your choice and have circled your answer, we would like to know how sure/confident you are that your answer is correct. Since there are three alternative answers and only one of them is correct you have a 33% chance of giving a correct answer. Therefore 33% means that you are guessing and do not know the correct answer, and 100% corresponds to absolute certainty.

You can use any number between 33% and 100% to indicate your confidence that your answer is correct.

Enter your confidence for every answer in the gap in the question after every test item:

How confident are you that your answer is correct? _____ %

Please answer all questions. Even if you have to guess everything, you could answer 33% correct by chance. You are not allowed to consult anyone else, or copy the answers from somebody.

NOTE: Please answer all questions, one after another in order in which they are presented in the questionnaire. Guess any answers you do not know. Do not jump around the questions, and do not return to already answered questions to change your answers; we are interested in your first answer.

You will be paid the money only if you have filled in the WHOLE questionnaire! Don't leave unanswered questions or unfilled gaps!

Please ask questions if something is unclear to you.

Thank you for your patience in completing this questionnaire.

Your personal data will be treated confidentially.

Surname, Name: _____

Gender: _____

Age:_____

Nationality:_____

Field of Study:_____

Semester:_____

Would you like to participate in another experiment, in which you can also win money?

Yes ☐ E-Mail: _____	No ☐

1.	How does one still call an instant camera? (circle one) Canon camera **Polaroid camera** Minolta camera
	How confident are you that your answer is correct? _____ %

2.	Where do flounders mostly live? (circle one) in coral reef **dug on the ground** in the reed
	How confident are you that your answer is correct? _____ %

3.	What does the *rollmops* consist of? (circle one) **herring filet** pork salmon filet
	How confident are you that your answer is correct? _____ %

4.	Which land does the Nobel Prize winner in Literature Gabriel García Márquez come from? (circle one) **Colombia** Spain Venezuela
	How confident are you that your answer is correct? _____ %

5.	Which style movement does anacreontics belong to? (circle one) **Rococo** Romanticism Realism
	How confident are you that your answer is correct? _____ %

6.	What is a hot chili sauce? (circle one) **Tabasco** Curacao Macao
	How confident are you that your answer is correct? _____ %

7.	How many letters does the Russian alphabet consist of? (circle one) 40 **33** 26
	How confident are you that your answer is correct? _____ %

8.	"Tosca" is an opera from ...? (circle one) **G. Puccini** G. Verdi A. Vivaldi
	How confident are you that your answer is correct? _____ %

9.	What s the name of the Greek Goddess of wisdom? (circle one) **Pallas Athena** Nike Penelope
	How confident are you that your answer is correct? _____ %

10.	Which is the most abundant metal on the Earth? (circle one) iron **aluminum** copper
	How confident are you that your answer is correct? _____ %

11.	How does one call an unknowing person? (circle one) Ignatius **ignorant** ideologue
	How confident are you that your answer is correct? _____ %

12.	Who flew for the first time with an airship around the Eiffel Tower? (circle one) **Santos-Dumont** count Zeppelin Saint-Exupéry
	How confident are you that your answer is correct? _____ %

13.	How is the snow shelter of Eskimos called? (circle one) wigwam **igloo** tipi
	How confident are you that your answer is correct? _____ %

14.	Which enterprise does Bill Gates belong to? (circle one) Intel **Microsoft** Dell Computers
	How confident are you that your answer is correct? _____ %

15.	How is the fasting month in Islam called? (circle one) Sharia **Ramadan** Imam
	How confident are you that your answer is correct? _____ %

16.	Which language does the concept "Fata Morgana" come from? (circle one) **Italian** Arabic Swahili
	How confident are you that your answer is correct? _____ %

17.	How many days does a hen need to incubate an egg? (circle one) **21 day** 14 days 28 days
	How confident are you that your answer is correct? _____ %

18.	What is ascorbic acid? (circle one) apple vinegar **vitamin C** vitamin A
	How confident are you that your answer is correct? _____ %