Thoughts on quantifying overconfidence in economic experiments

Michailova, Julija and Katter, Joana K. Q.

Helmut-Schmidt Univeristy, York Univeristy

January 2013

Thoughts on quantifying overconfidence in economic experiments

Julija Michailova[1]
Department of Economics, Helmut Schmidt University, Holstenhofweg 85,
22043 Hamburg, Germany
Phone: +494065413401
Fax: +494065412043
E-mail: julija_michailova@yahoo.com

Joana K. Q. Katter
Department of Psychology, York University, 309 Behavioural Science
Bldg., 4700 Keele St., Toronto, Canada M3J1P3
E-mail: katterjo@yorku.ca

Abstract

This article illustrates the difficulties in quantifying overconfidence in experimental finance and outlines a procedure for the development of a reliable overconfidence measurement instrument. Following the suggested two-stage procedure a sample measure of overconfidence is developed. First a pilot test is conducted to divide the initial fifty items into three difficulty levels: hard, moderate and easy questions. A final test was compiled of six questions of each difficulty levels. In the second phase a replicability check was run with the final instrument.

Keywords: miscalibration; overconfidence; overconfidence quantification; instrument development; experimental finance; economic experiments.

[1] Corresponding author.

1. Introduction

A growing body of financial literature presents results of experiments on overconfidence. In the experimental finance literature the term overconfidence refers to a group of effects, including miscalibration, the better than average effect, and illusion of control. Moore and Healy (2008) define these effects subsequently as overprecision, overplacement and overestimation. The concept of miscalibration is based on evidence from cognitive psychological research which suggests that human beings overestimate the precision of their knowledge (cf. Lichtenstein et al. 1982). The inclination of people to exaggerate their talents is referred to as the better than average effect (cf. Taylor and Brown 1988). For example, in an early finding, Svenson (1981) determined that 82 percent of participants rank themselves as being among the 30 percent of drivers with the highest driving safety. Illusion of control is linked to the exaggeration of the degree to which one can control one's own fate (cf. Langer 1975). This effect can arise, from a sense of optimistic overconfidence, which represents overestimation of the probabilities of the events that are advantageous to the subject (cf. Griffin and Brenner, 2004).

The current paper uses the term overconfidence in its original psychological sense of miscalibration. Individual calibration is tested by comparing the percentage of questions that a participant has answered correctly with her average confidence in the answers. A person is considered to be well calibrated if the following condition is satisfied: over the long run of those responses made with confidence P, about P% are correct (Adams, 1957). However most people are not well-calibrated and demonstrate overconfidence, which manifests itself through a systematic deviation from perfect calibration that is defined as an "unwarranted belief in the correctness of one's answer" (Lichtenstein, Fischhoff and Phillips, 1977). When overconfidence is measured through confidence intervals participants' probability distributions are generally too tight (Lichtenstein, Fischhoff, and Phillips, 1982). That is, when a subject is asked to define an interval so that she is X% sure it will contain an unknown numerical value the proportion of real answers falling inside the interval is lower than X%. In a study by Alpert and Raiffa (1982) they found that the 50% certainty intervals provided by participants included the true quantity only 30% of the time, while 98% confidence intervals included the true quantity only about 60% of the time.

Theoretical models of overconfidence in financial markets have linked market overconfidence to occurrence of speculative bubbles (Scheinkman and Xiong, 2003), excessive trade (De Bondt and Thaler, 1985) and price volatility (Benos, 1998). Experimental studies testing for these theoretical assumptions are becoming increasingly common in the literature (cf. Kirchler and Maciejovsky, 2002; Deaves, Lüders and Luo, 2009; Michailova and Schmidt, 2011). However in experimental finance literature little theoretical attention has been paid to the construct and no conventional method for quantifying overconfidence has been developed. Some papers do not even attempt to present the numerical measurement of the degree of overconfidence and use various proxies instead (e.g. Barber and Odean, 2001; Statman et al., 2006; Kliger and Levy, 2010). Hereby, there is a danger that in the area of experimental finance overconfidence may be inadequately measured and influenced by other factors than the imperfection of human nature, e.g. by the inappropriateness of the task inasmuch as it might be unclear to subjects and lack motivation for active participation (see Fischhoff, 1982). The lack of a reliable overconfidence measure calls in to question the conclusions drawn by existing experimental research on overconfidence.

The current paper is aimed at drawing attention to the importance of overconfidence measurement in experimental finance, highlighting the theoretical considerations that need to be taken into account when developing a suitable measurement method, including issues of cultural bias and question difficulty. We will then illustrate these theoretical considerations with a practical example, developing and testing a sample measure of overconfidence.

This article proceeds as follows. Section 2 reports on current methods that are being used in financial literature to measure overconfidence. Section 3 outlines the necessary theoretical and practical considerations that need to be taken into account when constructing a measure of overconfidence. Section 4 presents the methodology of the test construction. Section 5 elaborates on the statistical data analysis. Section 6 analyzes the findings from the replicability check, and, finally Section 7 concludes.

2. Overconfidence measurement in financial literature

Interest on the topic of the economic consequences of individual and market overconfidence has generated a large body of empirical and experimental research. Rather than providing a direct measure of overconfidence, some of these studies only assess indirect measures of overconfidence. For example, in a paper examining the interdependence between overconfidence and high trading volume in the American stock market, Statman et al. (2006)

use high past returns as a proxy for the degree of overconfidence. They argue that after high past returns, the posterior volume of trade will be higher, as the history of successful investment increases the degree of overconfidence. The same proxy for overconfidence was used by Kim and Nofsinger (2007) for the Japanese stock market. Barber and Odean (2001) use the trader's gender as a proxy for overconfidence, based on their assumption that according to the psychological literature, women are less overconfident than men. In another paper Barber and Odean (2002) employ as a proxy of overconfidence changes in the trading patterns of the investors who switched from phone-based to online trading. Consistent with their hypotheses, they find that overconfidence is associated with excessive trading in online investors. Papers that use proxies for overconfidence do not allow for the numerical measurement of the degree of overconfidence, neither on the individual nor on the aggregated market level.

Some current research does utilize direct measures of overconfidence, permitting the construction of the overconfidence measure for each individual using general and/ or financial knowledge tests or forecasting tasks. To asses individual overconfidence (miscalibration) the most common procedure is confidence intervals estimation. For example, Kirchler and Maciejovsky (2002) investigated overconfidence within the context of an experimental asset market by measuring the miscalibration of subjects is measured before each trading period, via two price prediction tasks: point prediction with the confidence in forecast and 98% confidence interval prediction. Russo and Schoemaker (1992) developed a scale to measure the degree of overconfidence that has been used to examine the link of overconfidence to both trading performance (e.g. Biais et al., 2005) and trading activity (e.g. Deaves et al., 2004). Their test consisted of a number of general-knowledge questions (ten items in Biais et al., 2005; 20 items in Deaves et al., 2009) with known numerical answers for which subjects had to state 90% confidence intervals.

Real stock market data predictions can also be used to measure overconfidence. For example, Glaser and Weber (2007) explored the connection between overconfidence and individual trading volume of a sample of individual investors with online broker accounts. To measure miscalibration subjects were asked to state upper and lower bounds of 90% confidence interval to the five stock price predictions and five economy-related questions. Menkhoff et al. (2006) surveyed 117 fund managers in order to detect an impact of experience on overconfidence, risk taking and herding behaviour. In their study miscalibration is measured as a 90% confidence estimation of the DAX index forecast.

3. Theoretical considerations in constructing an overconfidence measure

There are several theoretical considerations that need to be taken into account when constructing a measure of overconfidence. Overconfidence is most simply measured using knowledge tests. However, the degree of overconfidence is connected to the complexity of the task, such that overconfidence will be most pronounced for hard questions (few people know the correct answer) and least pronounced for easy ones (most people know the answer). An instrument that does not take the hard-easy effect into account in balancing the question difficulty may be prone to floor or ceiling effects of overconfidence (cf. Gigerenzer, Hoffrage and Kleinbölting, 1991). Proper calibration requires an initial administration of a large pool of items so that group accuracy can be estimated. The test items can then be drawn from this pool, where difficulty is determined based on the number of correct responses in the pilot testing. The individual overconfidence measures can then be balanced to the hard-easy effect by the inclusion of an equal number of questions from three difficulty levels (hard, medium and easy).

A related consideration is the potential for a cultural bias in the selected items, (cf. van Hemert, Baerveldt and Vermande, 2001) which occurs when the general knowledge questions are culture specific and might be familiar (i.e. easy) for individuals from a certain cultural or geographic group, but unfamiliar (i.e. hard) for individuals from another. A similar problem may occur with regards to a gender bias or a bias based on socioeconomic status. In order to avoid giving any "group of participants a relative advantage because of subject content" (Deaves et al., 2009), the target group for which the questionnaire is intended needs to be clearly defined and the items selected from the general knowledge domain of the target population. Particular care should be taken to ensure that the items are gender neutral. A sample drawn from the target group is then needed to test and validate the measure (Kennedy, Tarnai and Wolf, 2010).

Another set of considerations relate to the quantitative properties of the test. There are two types of calibration assessment techniques used in the psychological experiments: making probability judgments about discrete propositions and calibration of probability density functions assessed for numerical quantities (interval elicitation). With regards to test format, most previous research used confidence interval elicitation tasks to assess overconfidence (see Russo and Schoemaker, 1992). To measure overconfidence with this method, the assessor has to state for a series of questions with known numerical answer, upper and lower limits such that she is X% sure that the real answer would fall into that interval (cf. Lichtenstein et al., 1982). This method produces extreme overconfidence levels (cf. Juslin et al., 1999; Klayman et al., 1999; Winman et al., 2004).[1]

A better question format is to use discrete propositions with multiple-choice alternatives. To measure overconfidence with discrete propositions subjects are suggested to answer a series of questions and state their confidence for every question that their answer was correct (cf. Lichtenstein et al., 1982; Gigerenzer et al., 1991). These tasks are clearer to subjects and are not inherently prone to extreme overconfidence levels (cf. Klayman et al., 1999). Also, overconfidence in experimental financial literature is often assessed based on the insufficient number of items, e.g. Menkhoff, Schmidt and Brozynski (2006) used three assignments and Barber and Odean (2002) only two assignments to measure individual overconfidence. Recommendations from psychometrics suggest a minimum of 10-items to be used to provide a stable measure of a construct (cf. Kline, 1993).

Last but not least, overconfidence measurement should be administered with supervision, and should involve a financial incentive. This helps to reduce the desire of subjects to share the answers and increase the precision of the obtained individual bias scores. 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 questions.

3. Method

Procedure and subjects

A pilot study was conducted using 50 social science students from the Christian-Albrechts University of Kiel. Participants were instructed to fill in the 50-question test (*test-50*) in approximately 30 minutes (instructions are available in Appendix A). Three monetary prizes were offered to those participants who got the most items right. Twenty five of the participants were male and 25 female. They had a mean age of 24.32 (SD = 0.31) and have studied from 3 to 11 semesters (M = 6.98, SD = 2.11). Most of the subjects were German (94.8%).

Design and materials

For the pilot test 50 general knowledge questions were selected from the German quiz web-page http://wissen.de. Each question had three short (one or two-word) multiple choice answers. Students had to answer all the questions and state their 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. Individual overconfidence was measured as a bias score, which was calculated as the difference between the average confidence level across all questions and the proportion of correct answers (see Equation 1). A positive score represented overconfidence, a negative score represented underconfidence, and a bias score of zero indicated an accurately calibrated person.

$$bias\ score = \frac{1}{N}\sum_{i=1}^{N}\left(c_i - a_i\right) \qquad (1)$$

where N is the number of test items; c_i is the confidence in answering item i; a_i is the accuracy in answering item i (takes value 100 if true, or 0 otherwise). In our example the minimum value of the bias score is -67 (a person has answered all questions right, but was completely unsure in his answers) and the maximum value is 100 (a person has answered all questions wrong, but was completely sure his answers were right).

In addition to measuring how well the subjects were calibrated, some personal data were collected: name, age, educational background, duration of studies in semesters, and nationality. At the beginning of the pilot session 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.

Based on the analysis of the pilot-test outcomes, a final test (*test-18*) was constructed from 18 questions of the three difficulty levels: six hard, six medium and six easy questions (Table 1). Items were assigned to the three difficulty levels based on the average group accuracy: 0-33% accuracy hard questions, 34-66% medium difficulty, 67-100% easy questions. After the initial division, four questions have fallen in the category of hard questions (average accuracy 17.5%), 10 in the category of medium questions (average accuracy 55.2%) and 36 in the category of easy questions (average accuracy 88.5%). Since there were not enough hard questions, based on the idea that overconfidence is the most pronounced for hard questions (cf. Pitz, 1974), average overconfidence ratio over each of the medium questions was calculated and two with the highest value were added to hard questions.

Insert Table 1 about here

4. Results

Consistent with the previous research, on average subjects were 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* increased the average overconfidence measure (M = 14.11, SD = 10.63). Table 2 presents the bias scores and accuracy of all participants of the pilot study for both *test-50* and *test-18*, and males and females separately.

Insert Table 2 about here

It shall be noted that whereas for the *test-50* average overconfidence of men was slightly lower than that of women, after recalculating the overconfidence ratio for the *test-18*, the average bias score for both groups became almost identical. Pearson's correlation analysis has not detected any significant linear relationship between the individual bias scores and individual age (*test-50*: Pearson correlation(48) = -0.629, p = 0.377, one-sided; *test-18*: Pearson correlation(48) = 0.078, p = 0.312, one-sided) or duration of study in semesters (*test-50*: Pearson correlation(48) = 0.148, p = 0.152, one-sided; *test-18*: Pearson correlation(48) = 0.194, p = 0.088, one-sided). Thus students of different age groups and being at different levels of progress with their studies can be recruited for participation at financial overconfidence experiments. 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). This is in line with previous findings (cf. Brenner et al., 1996)

Test-50 versus test-18: accuracy and confidence

Analysis of the group accuracy for *test-50* revealed that 72% of the questions were easy (67-100% accuracy) (see Figure 1 (a). This test was characterized by high precision and low confidence, consequently 58% of the questions resulted in average underconfidence (Figure 1(b). This outcome illustrates the danger of using the unbalanced to hard-easy effect test for quantifying individual overconfidence.

Insert Figure 1 about here

Employment of *test-50* would result in average group underconfidence[2], as 24% of subjects who completed *test-50* were underconfident (see Figure 2(a); for *test-18* this number decreased to 8% (see Figure 2(b). Comparison of the *test-50* to *test-18*, revealed that the later also results in the improvement in the symmetry of the distribution of the bias score (*test-50*: skewness = 0.73; *test-18*: skewness = 0.53). Alongside an increase in the range of the bias score is observed (from 38.60 for *test-50* to 47.23 for *test-18*). This increase is important for experimental studies as it leaves more room for finding subjects, whose degree of overconfidence differs significantly, thus allowing testing hypotheses about the influence of individual degree of overconfidence on experimental outcomes.

Insert Figure 2 about here

Statistical Tests

This section presents the results of the statistical tests that verify the success of the categorization of the questions into three levels of difficulty for the *test-18*. Characteristics of the final test in terms of the confidence, accuracy and the bias score are presented in the Table 3.

Insert Table 3 about here

Participants exhibited overconfidence for hard and medium questions, and underconfidence for easy questions. This is in line with previous research, which found hard questions to be the most prone to overconfidence and easy questions often to be subject to underconfidence (e.g. Pitz, 1974; Lichtenstein et al., 1982). 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 medium 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 medium questions produced in average the bias score which was the most indistinguishable from the perfect calibration score of zero. Kruskal-Wallis H Test indicated that the three difficulty levels of questions resulted in significantly different levels of accuracy (Chi-squared (2) = 15.760, p = 0.00; effect size η^2 = 0.926), confidence (Chi-squared (2) = 11.617, $p < 0.01$; effect size η^2 = 0.856) and bias scores (Chi-squared (2) = 12.117, $p < 0.01$; effect size η^2 = 0.783). These results proved that the division of questions into three difficulty levels was successful.

Gender Differences *Test-50:* No statistically significant difference was found in overconfidence between the two genders (t(48) = -1.109, p = 0.27, two-sided). However, males were significantly more accurate (t(48) = 3.053, $p < 0.01$, one-sided; effect size η^2 = 0.163) and confident than females (t(48) = 1.840, $p < 0.05$, one-sided; effect size η^2 = 0.069), which suggests a gender bias in the pilot test items. 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). *Test-18:* Difference between males and females in confidence (t(48) = 1.37, p = 0.176, two-sided; effect size η^2 = 0.037), accuracy (t(48) = 0.704, p = 0.485, two-sided; effect size η^2 = 0.01) and overconfidence (t(48) = -0.002, p = 0.998, two-sided; effect size η^2 = 0.00) was insignificant. No significant difference in overconfidence was found between male and female subjects for the three levels of question difficulty (hard: t(48) = 0.085, p = 0.933, two-sided; medium: t(48) = 0.354, p = 0.725, two-sided; easy: 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) = -0.630, $p < 0.01$, one-sided; women: Pearson's Correlation (23) = -0.625, $p < 0.01$, one-sided).

5. Replicability check

The study was repeated with different students from the target group. A total of 34 participants, 21 males and 13 females, were given approximately 15 minutes to fill in the final overconfidence test (*test-18*). As in the pilot study, three monetary prizes were offered to the subjects who got the most items right. Participants had a mean age of 26.06 (SD =

2.62) and have on average studied 9.10 semesters (SD = 2.60). Most of the subjects were German (86%). On average subjects were found to be overconfident (M = 10.41, SD = 9.26). Average group overconfidence in *test-18* did not significantly differ between the pilot and the replicability check (t(82) = 1.649, p = 0.103, two-sided; size effect $\eta^2 = 0.032$). Table 4 presents the bias scores and accuracy of all participants of the study, and males and females separately. As in the pilot study, no significant linear relationship between the individual bias scores and individual age (Pearson correlation(32) = 0.189, p = 0.142, one-sided) or duration of study in semesters (Pearson correlation(32) = -0.054, p = 0.338, one-sided) could be detected._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).

Insert Table 4 about here

Gender differences Difference between male and female participants in confidence (t(32) = -0.53, p = 0.600, two-sided; effect size $\eta^2 = 0.009$), accuracy (t(32) = -0.524, p = 0.604, two-sided; effect size $\eta^2 = 0.009$) and overconfidence (t(32) = 0.211, p = 0.834, two-sided; effect size $\eta^2 = 0.001$) were insignificant. No significant difference in overconfidence was found between male and female subjects for the three levels of question difficulty (hard: t(32) = 0.042, p = 0.967, two-sided; medium: t(32) = -0.357, p = 0.723, two-sided; easy: t(32) = 1.468, p = 0.152, two-sided).

Reliability DeCoster (2000, p. 1) notes that a scale can be called reliable “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. For the instrument two values of Cronbach’s alpha were calculated: $\alpha_{\text{confidence}} = 0.79$ and $\alpha_{\text{overconfidence}} = 0.68$. Values of the calculated alphas were either close or exceeded the threshold values, considered optimal for the use in social research (cf. Moss et al., 1993). Thus, the developed instrument possesses good internal consistency (reliability).

6. Conclusions

This article demonstrates the difficulties of quantifying overconfidence in experimental finance and suggests a procedure for the development of the reliable overconfidence measure. The principal steps to improve the instrument were: 1) choice of another test format (discrete propositions with multiple-choice alternatives instead of confidence intervals), 2) balancing the test for the hard-easy effect, and 3) controlling for gender and country bias. Following the suggested procedure an instrument is developed in a two-stage procedure. In the first phase a pilot test was conducted to assess questions' difficulty, based on the group accuracy in answering the initial test items. Subsequently, six questions of the three difficulty types were included in the final test. The second phase was aimed at verification of replicability of results. Both studies were administered with the students of the target group, who were offered a reward on the basis of competition in test accuracy. Evidence was found for the significant effect of the question difficulty on the overconfidence measure and for the existence of the gender bias. The statistical analysis confirmed that the three types of questions significantly differed from each other in terms of the produced confidence, accuracy and overconfidence. In the created instrument gender is not associated with overconfidence. The instrument's reliability is acceptable for the use in social research. 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 experimental studies, it can be concluded that the instrument suitable for evaluation of individual differences in the degree overconfidence was created.

Acknowledgements

We want to express gratitude to Briony D. Pulford for a fruitful cooperation. We thank Jürgen Golz for valuable comments. A special thank to Christian Radden, who helped in translating the experimental instructions to German. Julija Michailova acknowledges a German Academic Exchange Office (DAAD) scholarship.

References

Adams, J. K. (1957) ‘A confidence scale defined in terms of expected percentages.’ *The American Journal of Psychology*, Vol. 70, pp. 432–436.

Alpert, M., and Raiffa, H. (1982) ‘A progress report on the training of probability assessors’, in Kahneman, D. et al (Eds.), *Judgment under Uncertainty: Heuristics and Biases*, Cambridge University Press, New York, pp. 294–305.

Barber, B., and Odean, T. (2001) ‘Boys will be boys: Gender, overconfidence and common stocks investments.’ *Quarterly Journal of Economics*, Vol. 116, pp. 261–292.

Barber, B., and Odean, T. 2002) ‘Online investors: do the slow die first?’ *Review of Financial Studies*, Vol. 15, pp. 455–487.

Benos, A. (1998) ‘Aggressiveness and Survival of Overconfident Traders.’ *Journal of Financial Markets*, Vol. 1, pp. 353–383.

Brenner, L. A., Koehler, D. J., Liberman, V. and Tversky, A. (1996) ‘Overconfidence in probability and frequency judgments: A critical examination.’ *Organizational Behavior and Human Decision Processes*, Vol. 65, No. 3, pp. 212–219.

Deaves, R., Lüders, E. and Luo, G. Y. (2009) ‘An experimental test of the impact of overconfidence and gender on trading activity.’ *Review of Finance*, Vol. 13, pp. 555–575.

De Bondt, W. F. M., and Thaler, R. (1985) ‘Does the Stock Market Overreact?” *The Journal of Finance*, Vol. 40, pp. 793–805.

DeCoster, J. (2005) ‘Scale Construction Notes.’ Retrieved <04.30.2013> from http://www.stat-help.com/notes.html

Fischhoff, B., (1982) ‘Debiasing’, in Kahneman, D. et al (Eds.), *Judgment under Uncertainty: Heuristics and Biases*, Cambridge University Press, New York, pp. 422–444.

Gigerenzer, G., Hoffrage, U. and Kleinbölting, H. (1991) ‘Probabilistic mental models: A Brunswikian theory of confidence.’ *Psychological Review*, Vol. 98, pp. 506–28.

Glaser, M., and Weber, M. (2007) ‘Overconfidence and trading volume.’ *The Geneva Risk and Insurance Review*, Vol. 32, pp. 1–36.

Griffin, D., and Brenner, L. (2004) ‘Perspectives on probability judgment calibration’, in Koehler, D. and Harvey, N. (Eds.), *Blackwell Handbook of Judgment and Decision Making*, Blackwell, Malden Mass, pp. 177–199.

Juslin, P., Wennerholm, P. and Olsson, H. (1999) ‘Format dependency in subjective probability calibration.’ *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 28, pp. 1038–1052.

Juslin, P., Winman, A. and Hansson, P. (2007) ‘The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals”. *Psychological Review*, Vol. 114, pp. 678–703.

Kennedy, J. M., Tarnai, J. and Wolf, J. G. (2010) ‘Managing survey research projects’, in Marsden, P. V. and Wright, J. D. (Eds.), *Handbook of survey research*, Emerald Group Publishing Ltd., Bingley, pp. 575–592.

Kim, A. K., and Nofsinger, J. R. (2003) ‘The behavior and performance of individual investors in Japan.’ *Pacific Basin Finance Journal*, Vol. 11, pp. 1–22.

Kliger, D., and Levy, O. (2010) ‘Overconfident investors and probability misjudgements.’ *The Journal of Socio-Economics*, Vol. 39, pp. 24–29.

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

Kline, P., (1993) *The Handbook of Psychological Testing*, New York, London, Routledge.

Kirchler, E., and Maciejovsky, B. (2002) ‘Simultaneous over- and underconfidence: Evidence from Experimental Asset Markets.’ *Journal of Risk and Uncertainty*, Vol. 25, pp. 65–85.

Langer, E., (1975) ‘The illusion of control.’ *Journal of Personality and Social Psychology*, Vol. 32, pp. 311–328.

Lichtenstein, S., Fischhoff, B. and Phillips, L. D. (1977) ‘Calibration of probabilities: The state of the art’, in Jungermann, H. and deZeeuw, G. (Eds.), *Decision Making and Change in Human Affairs*, D. Reidel, Amsterdam, pp. 275–324.

Lichtenstein, S., Fischhoff, B. and Phillips, L. D. (1982) 'Calibration of probabilities: the state of the art to 1980', in Kahneman, D. et al (Eds.), *Judgment under Uncertainty: Heuristics and Biases*, Cambridge University Press, New York, pp. 306–334.

Michailova, J., and Schmidt, U. (2011) 'Overconfidence and bubbles in experimental asset markets.' Kiel Working Papers 1729, Kiel Institute for the World Economy.

Menkhoff, L., Schmidt U. and Brozynski, T. (2006) 'The impact of experience on risk taking, overconfidence, and herding of fund managers: Complementary survey evidence.' *European Economic Review*, Vol. 50, pp. 1753–1766.

Moore, D.A, and Healy, P. J. (2008) 'The trouble with overconfidence.' *Psychological Review*, Vol. 115, pp. 502–517.

Moss, S., Patel, P., Prosser, Goldber, H. D., Simpson, N., Rowe, S. and 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, pp. 471–480.

Pitz, G.F., (1974) 'Subjective probability distributions for imperfectly known quantities', in Gregg, L. W. (Ed.), *Knowledge and Cognition*, Wiley, New York, pp. 29–41.

Russo, J. E., and Schoemaker, P. J. (1992) 'Managing overconfidence.' *Sloan Management Review*, Vol. 33, pp. 7–17.

Scheinkman, J. A., and Xiong, W. (2003) 'Overconfidence and Speculative Bubbles.' *Journal of Political Economy*, Vol. 111, pp. 1183–1219.

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

Svenson, O., (1981) 'Are we all less risky and more skilful than our fellow drivers?" *Acta Psychologica*, Vol. 47, No. 2, pp. 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, pp. 193–210.

van Hemert, D. A., Baerveldt, C., and Vermande, M. (2001) ‘Assessing cross-cultural item bias in questionnaires. Acculturation and the measurement of social support and family cohesion for adolescents.’ *Journal of Cross-Cultural Psychology*, Vol. 32, pp. 381–396.

Winman, A., Hansson, P. and Juslin, P. (2004) ‘Subjective probability intervals: How to reduce overconfidence by interval evaluation.’ *Journal of Experimental Psychology: Learning Memory and Cognition*, Vol. *30*, pp. 1167–1175.

Endnotes

1. Articles by Winman, Hansson and Juslin (2004) and Juslin, Winman and Hansson (2007) are analyzing the possible reasons of extreme overconfidence production by confidence (probability) intervals.

2. A test skewed in the direction of hard questions would result in group overconfidence

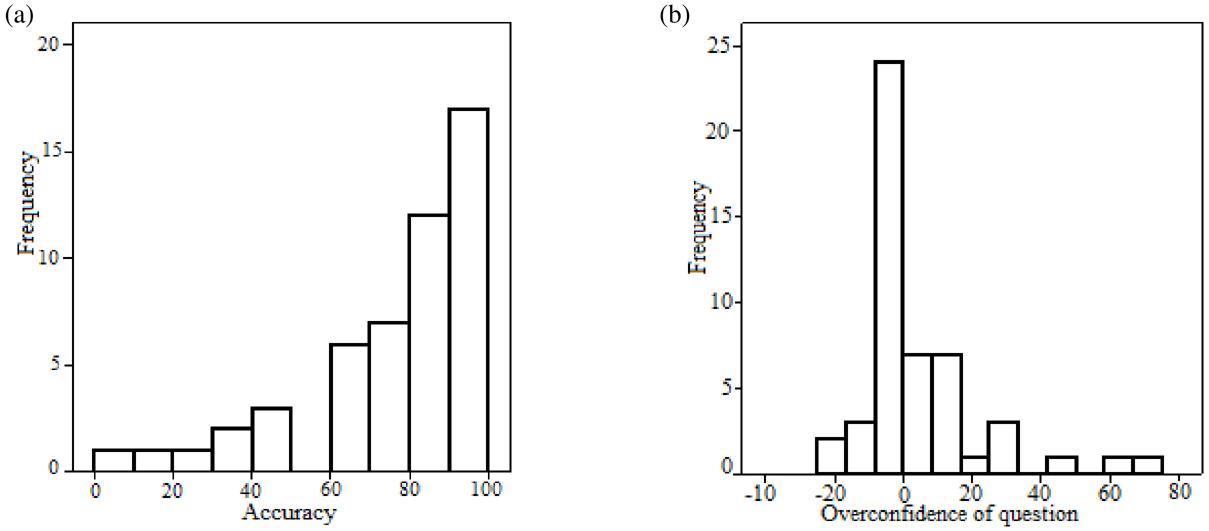

Figure1. Distribution of accuracy (a) and overconfidence per question (b) in *test-50*

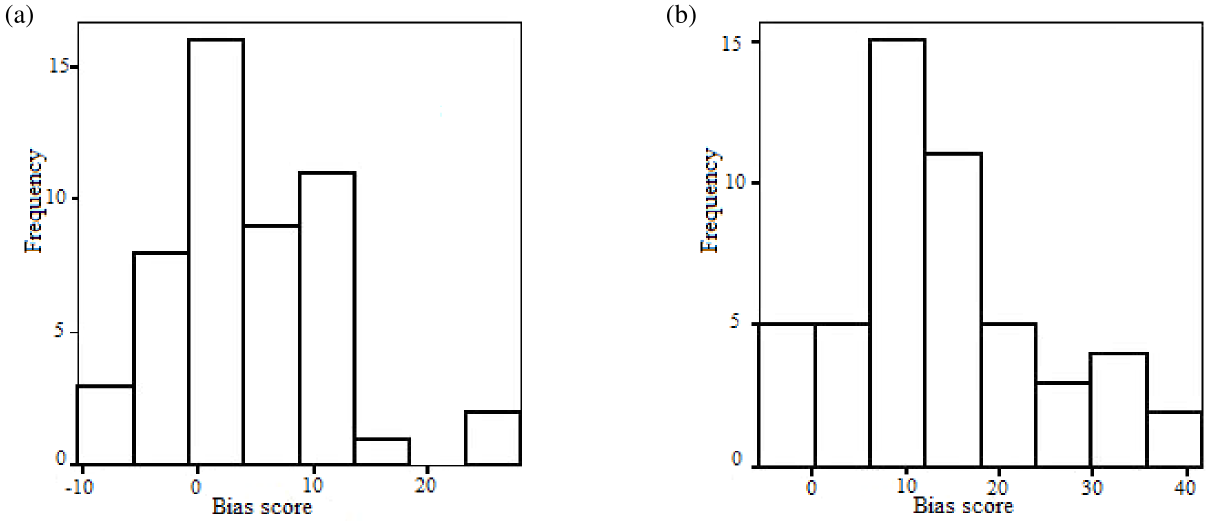

Figure 2. Distribution of individual bias scores by test: (a) *test-50*; (b) *test-18*

Table 1

Items included in *test-18* (translation from German; arranged from easy to hard; correct answer is underlined)

What is the name for an instant camera?	Canon camera	Polaroid camera	Minolta camera
What is a rollmop made of?	herring	pork	salmon
What is a hot chilli sauce?	Tabasco	Curacao	Macao
What is the name of Eskimo snow shelter?	wigwam	igloo	tipi
What enterprise belongs to Bill Gates?	Intel	Microsoft	Dell Computers
What is the Islamic month of fasting called?	Sharia	Ramadan	Imam
Where do flounders usually live?	among the reeds	amongst coral reefs	on the sea bottom
What country does the Nobel Prize winner in Literature Gabriel García Márquez come from?	Spain	Venezuela	Colombia
What artistic movement does anacreontics belong to?	Rococo	Romanticism	Realism
How many letters are there in the Russian alphabet?	40	33	26
What is the name of the Greek Goddess of wisdom?	Pallas Athena	Nike	Penelope
What is ascorbic acid?	apple vinegar	vitamin A	vitamin C
"Tosca" is an opera by ...?	G. Puccini	G. Verdi	A. Vivaldi
What is the most abundant metal on Earth?	iron	aluminium	copper
What is a word to describe an unknowing person?	Ignatius	ignorant	ideologue
Who was the first person to fly around the Eiffel Tower in an airship?	Santos-Dumont	count Zeppelin	Saint-Exupéry
What language does the term "Fata Morgana" come from?	Arabic	Swahili	Italian
How long does it take for a hen to hatch an egg?	21 days	14 days	28 days

Table 2

Pilot study: Overconfidence and accuracy

Test 50

		Overconfidence				Accuracy			
OBS	Group	Mean	SD	Min	Max	Mean	SD	Min	Max
50	All	4.48	7.34	-10.40	28.20	76.16	6.61	58	90
25	Female	5.63	8.47	-8.40	28.20	73.52	6.72	58	84
25	Male	3.33	5.96	-10.40	13.00	78.80	5.45	66	90

Test 18

		Overconfidence				Accuracy			
OBS	Group	Mean	SD	Min	Max	Mean	SD	Min	Max
50	All	14.11	10.63	-5.56	41.67	62.78	9.99	38.89	83.33
25	Female	14.12	10.79	-5.56	41.67	61.78	10.43	38.89	77.78
25	Male	14.11	10.70	-3.89	36.11	63.78	9.64	44.44	83.33

Table 3

Pilot study: Numerical characteristics of the *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

Table 4

Replicability check: Overconfidence and accuracy

		Overconfidence				Accuracy			
OBS	Group	Mean	SD	Min	Max	Mean	SD	Min	Max
34	All	10.41	9.26	-6.28	30.00	60.46	9.35	38.89	77.78
13	Female	9.98	8.68	-3.44	28.94	61.54	9.48	38.89	77.78
21	Male	10.68	9.81	-6.28	30.00	59.79	9.45	38.89	77.78

Appendix A

Experimental Instructions (text is based on the sample received from Dr. Briony D. Pulford)

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:____

1.	What is the name for an instant camera? (circle one) Canon camera **Polaroid camera** Minolta camera
	How confident are you that your answer is correct? _______ %

2.	Where do flounders usually live? (circle one) among the reeds amongst coral reefs **on the sea bottom**
	How confident are you that your answer is correct? _______ %

3.	What is a rollmop made of? (circle one) **herring** pork salmon
	How confident are you that your answer is correct? _______ %

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

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

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

7.	How many letters are there in the Russian alphabet? (circle one) 40 **33** 26
	How confident are you that your answer is correct? _______ %

8.	“Tosca” is an opera by ...? (circle one) **G. Puccini** G. Verdi A. Vivaldi
	How confident are you that your answer is correct? _______ %

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

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

11.	What is a word to describe an unknowing person? (circle one) Ignatius **ignorant** ideologue
	How confident are you that your answer is correct? _______ %

12.	Who was the first person to fly around the Eiffel Tower in an airship? (circle one) **Santos-Dumont** count Zeppelin Saint-Exupéry
	How confident are you that your answer is correct? _______ %

13.	What is the name of Eskimo snow shelter? (circle one) wigwam **igloo** tipi
	How confident are you that your answer is correct? _______ %

14.	What enterprise belongs to Bill Gates? (circle one) Intel **Microsoft** Dell Computers
	How confident are you that your answer is correct? _______ %

15.	What is the Islamic month of fasting called? (circle one) Sharia **Ramadan** Imam
	How confident are you that your answer is correct? _______ %

17.	How long does it take for a hen to hatch an egg? (circle one) **21 days** 14 days 28 days
	How confident are you that your answer is correct? _______ %

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