# Confidence Interval Estimation Tasks and the Economics of Overconfidence

David Cesarini*, Örjan Sandewall**, Magnus Johannesson***

SSE/EFI Working Paper Series in Economics and Finance

No 535

September 2003

## Abstract

Experiments in psychology, where subjects estimate confidence intervals to a series of factual questions, have shown that individuals report far too narrow intervals. This has been interpreted as evidence of overconfidence in the preciseness of knowledge, a potentially serious violation of the rationality assumption in economics. Following these results a growing literature in economics has incorporated overconfidence in models of, for instance, financial markets. In this paper we investigate the robustness of results from confidence interval estimation tasks with respect to a number of manipulations: frequency assessments, peer frequency assessments, iteration, and monetary incentives. Our results suggest that a large share of the overconfidence in interval estimation tasks is an artifact of the response format. Using frequencies and monetary incentives reduces the measured overconfidence in the confidence interval method by about 65%. The results are consistent with the notion that subjects have a deep aversion to setting broad confidence intervals, a reluctance that we attribute to a socially rational trade-off between informativeness and accuracy.

JEL Classification: C91, D80, Z13.

Key words: overconfidence, uncertainty, monetary incentives, experiments.

* Address: Department of Economics, London School of Economics, Houghton Street London WC2A 2AE, United Kingdom; e-mail: d.a.cesarini@lse.ac.uk

** Address: Department of Economics, London School of Economics, Houghton Street London WC2A 2AE, United Kingdom; e-mail: n.o.sandewall@lse.ac.uk

*** Address: Department of Economics, Stockholm School of Economics, Box 6501, SE-113 83 Stockholm, Sweden; e-mail: hemj@hhs.se

Acknowledgements: The authors thank J. Edward Russo for generously taking the time to engage in helpful e-mail exchange. This paper has also benefited from discussions with Christopher Brooke, Cesar Calvo, Erika Färnstrand and Tino Sanandaji. Financial support from the Jan Wallander and Tom Hedelius Foundation is gratefully acknowledged.

# 1. Introduction

*'When Parisian taxi drivers want to press a point of the municipal authorities about regulations or fees, they sometimes launch a work-to-rule strike. It consists merely in following meticulously all the regulations in the Code routier and thereby bringing traffic throughout central Paris to a grinding halt. The drivers thus take tactical advantage of the fact that the circulation of traffic is possible only because drivers have mastered a set of practices that had evolved outside, and often in contravention, of the formal rules.'* (Scott, 1998, p. 256)

A growing literature in economics explores the economic consequences of overconfidence – a psychological bias often considered an important departure from the *homo oeconomicus* paradigm. The term overconfidence has been used to describe two distinct phenomena. The first is the tendency of individuals to express excessive belief in their own capacity, e.g. the ability to drive safely (Svenson, 1981). The second phenomenon is the tendency of individuals to overestimate the preciseness of their knowledge, i.e., a lack of metacognitive capacity.[1] Henceforth, we will use the term overconfidence to denote the second phenomenon.

Alpert and Raiffa (1969) are usually credited with the first 'discovery' of overconfidence. However, the most influential study is probably a paper by Russo and Schoemaker (1992) in the Sloan Management Review, which has often been taken as evidence that managers act overconfidently. This article has reached an academic audience outside the realm of management and psychology, and the Science Citations Index reports references to this paper in – amongst many others – the Quarterly Journal of Economics (Barber and Odean, 2001),

[1] The term overconfidence has furthermore been used to denote the tendency of people to express excessive optimism concerning the probability of a certain favorable/unfavorable outcome in the future (Babad, 1987).

the American Economic Review (Moskowitz and Vissing-Jorgensen, 2002) and the Journal of Finance (Kyle and Wang, 1997; Daniel et al., 1998; Odean, 1998).

Russo and Schoemaker (1992) use a confidence interval assessment test where the test subjects are given the following instructions: (Russo and Schoemaker, 1992, p. 8) 'For each of the following questions, provide a low and a high estimate such that you are 90 percent certain the correct answer will fall within these limits. You should aim to have 90 percent hits and 10 percent misses.' Ten tailor-made questions are then provided, in some cases at the industry level and in some cases at the firm specific level. The sample is roughly 2000 professionals. Even though a well-calibrated individual following the test instructions should on average err in 10 percent of cases if a confidence interval of 90 percent is provided, typical outcomes are in the range of 50–60 percent.[2]

The seemingly substantial overconfidence in the Russo and Schoemaker (1992) study has also been documented in many other interval estimation studies in psychology (see, for example, Juslin et al. (1999)), and the mainstream view in psychology is that it is a very important phenomenon. Myers (1993, p. 126) for example, refers to overconfidence as a 'fact of psychology' and von Winterfeldt and Edwards (1986, p. 539) call it a 'reliable and reproducible finding'. If overconfidence is a pervasive feature of behavior, this will also have profound implications for economics. Overconfidence will, for instance, affect behavior on financial markets. Recently, a number of theoretical models on financial markets that attempt to incorporate overconfident judgments have also been developed (Odean, 1998; De Long et al., 1991; Kyle and Wang, 1997). Odean (1999) provides evidence of excessive trading and negative abnormal returns amongst certain stock market traders and interprets this in terms of overconfidence. In the finance literature the presence of overconfidence has become well

[2] Other response formats than the confidence interval method have also been employed to measure overconfidence. For comprehensive reviews of the calibration literature see e.g. Keren, 1991; Lichtenstein et al., 1982; Yates, 1990.

established. DeLong et al. (1991, p. 10) for example refer to it as 'one of the best documented biases'.

Overconfidence may also be relevant for macroeconomics. The disagreements about the desirability of activist monetary policy originate in conflicting views about the preciseness with which policymakers can assess the contemporaneous state of the economy. In a similar vein, Orphanides (2000, p. 10) has recently argued that the cause of the great inflation was a reliance on the 'heroic assumption of perfect information regarding the state of the economy'. Due to the economic importance of the subject, it is important to scrutinize the empirical evidence on overconfidence. In this paper we question the economic relevance of the results of interval estimation studies on methodological grounds. Two fundamental distinctions are important for our argument: frequency assessments vis-à-vis subjective probability estimates and stated judgments vis-à-vis genuine judgments.

The interval estimation method is based on subjective probability estimates. An increasingly influential branch of calibration research (chief references include Cosmides and Tooby, 1996; Gigerenzer et al., 1991; Hoffrage et al., 2000) in psychology has established that there is a large and important distinction to be made between subjective probability estimates and assessments based on natural frequencies. For example, Tversky and Kahneman's famous discovery about the violation of the conjunction rule (Tversky and Kahneman, 1983) turned out to be considerably reduced in magnitude when test subjects were provided with information in frequency format (see Hertwig and Gigerenzer, 1999). Some authors (see e.g. Cosmides and Tooby, 1996) have extended this line of reasoning based on theories in evolutionary psychology, arguing that frequentist representations will normally elicit Bayesian, well-calibrated reasoning, thus challenging the heuristics and biases research program. Empirical and theoretical research thus indicates that human cognitive mechanisms are designed to operate on frequentist rather than probabilistic principles. If true, this

jeopardizes the validity of results obtained in tests using probabilistic input, such as in confidence interval estimation tasks.

Moreover, stated judgments should not necessarily be equated with actions. Whereas the experimental psychology literature on overconfidence is voluminous, surprisingly few attempts have been made to assess the behavioral consequences of overconfidence observed in a laboratory setting. Rather, it has been implicitly assumed that the stated judgments of test subjects are identical with 'genuine' judgments.[3] This may be problematic in interval estimation studies for two reasons. Firstly, no monetary incentives are provided to report true confidence intervals. In experimental economics it is standard practice to use monetary incentives, a reflection of a methodological difference between economics and psychology (Camerer and Hogarth, 1999; Hertwig and Ortmann, 2001). Secondly, subjects may not strictly follow the experimental instructions.

Previous research based on the social rationality paradigm suggests that subjects when asked to provide a subjective confidence interval will make a trade-off between accuracy and informativeness to adhere to social norms (Yaniv and Foster, 1995, 1997). For example, when asked a question, people will want to give an informative answer, as social conventions prescribe, and as they have been expected to do on numerous occasions (Grice, 1975). The maxim of relation - one of Grice's fundamental maxims of communication in his theory of conversational reasoning - states that people try to make an 'appropriate' contribution in each stage of communication (Grice, 1975). What then, can be regarded as an informative or appropriate answer? When someone asks for the time, he does not expect the answer in a 90 percent confidence interval, even if he is in great need of a correct answer. In general, an answer of the type 'I think it is roughly half past three', is more useful than 'my ninety

[3] We recognize the semantic difficulties inherent in the use of the word genuine. In this paper, we define a judgment as being 'genuine' if – and only if – it has behavioral consequences.

percent confidence interval is that the time is somewhere in between 1.30 and 5.15'. Even when the second answer encompasses the true value, its informational value is limited.

The claim that individuals make a trade-off between accuracy and informativeness is supported by experimental evidence. Yaniv and Foster (1995, 1997) in a series of experiments provided people with predetermined interval estimates of the number of United Nations member countries. The respondents were informed that the true value was 159 and were then asked to choose between two estimates: (A) '140–150' and (B) '50–300' (Yaniv and Foster, 1995). Approximately 90 percent favored the narrow interval. Yaniv and Foster interpret this result as saying that people are willing to sacrifice accuracy in return for informativeness when giving confidence intervals. If this is the case the answers to interval estimation tasks cannot be interpreted as evidence of overconfidence.

In this paper we experimentally test the stability of the results in interval estimation tasks. After providing ten 90% confidence intervals on factual questions, subjects are asked to estimate the number of correct answers (to compare frequency assessments vis-à-vis subjective probability). We also study the effects of iteration and financial incentives and investigate whether participants anticipate the overconfidence of others. According to our results using frequency assessments rather than confidence intervals dramatically reduce the measured overconfidence, although some overconfidence appears to remain. Providing monetary incentives in the frequency assessments further lowers measured overconfidence, but the effect of incentives is small and not significant. Subjects furthermore correctly anticipate that other subjects' confidence intervals are too narrow. We conclude that the commonly used confidence interval method substantially overestimates overconfidence. Just like French taxi drivers, subjects knowingly do not always follow the rules.

The design of the experiment and the hypotheses are described below. This is followed by a presentation of the results. We end with a discussion of the results and some conclusions.

# 2. Design of experiment and hypotheses

## 2.1 Subjects and procedures

A total of 85 undergraduates in Economics and Business (65 male, 20 female) at the Stockholm School of Economics participated in the experiment. They were recruited by advertisement and were paid a participation fee of SEK 50 (approximately USD 6 at the time of the experiment). The participants had prior knowledge in basic statistics and were familiar with the concept of a confidence interval. Subjects were randomly allocated between two experimental treatments: the control group (n=38) and the incentives group (n=47).[4] The only difference between the treatments is that in the incentives group subjects received an additional SEK 50 for each accurate reply in stages 2, 3, 5, and 6.[5] The six stages of the experiment are described below. The complete instructions are reported in the Appendix.

In stage 1 the participants were provided with ten numerical questions specific to the domain of economics (broadly interpreted). They were instructed to state a lower and an upper limit for each question such that their subjective confidence that the interval would actually contain the true value would be equal to ninety percent. The exact wording was taken from the Russo and Schoemaker (1992) study, with the addition of some clarifying remarks. The participants were informed that the questions had been randomly chosen from a pool of possible questions with no deliberate attempt to deceive or misguide them**.** The rationale for

[4] The groups were randomized by month of birth. The slight imbalance between number of participants in the incentives and the control group is purely coincidental, since our randomization procedure did not ensure an equal number of participants.

this representative sampling procedure was to address the alleged problem of oversampling of tricky questions that, according to some authors, have driven the overconfidence result in a number of studies (May, 1986). At no point were participants aware of how many stages of the experiment remained. We refer to this first stage as the interval assessment.

Following the first stage, the participants were asked to estimate how many of their own answers in Stage 1 that contained the true value. This was labeled Stage 2. In the third stage participants were asked how many questions they believed their peers had answered correctly (in the sense that the confidence interval contained the true value). We refer to these estimates in Stages 2 and 3 as the frequency assessment and the peer frequency assessment respectively.

Upon completion of the third stage, we instructed all subjects who had not provided a frequency assessment equal to nine to give new revised estimates to ensure that they captured the correct answer in ninety percent of cases. This was denoted Stage 4.[6] Following the completion of Stage 4, Stages 2 and 3 were repeated (now labeled Stages 5 and 6).

## 2.2 Hypotheses and tests

Let $\mu_1, \mu_2, ..., \mu_6$ refer to the population mean in each of the six stages in the control group and let $\mu_1^*, \mu_2^*, ..., \mu_6^*$ denote the corresponding means for the incentives group. Our hypotheses can be organized around the four manipulations of the interval method: frequency assessments, peer frequency assessments, iteration, and monetary incentives. We used an independent

---

5 An accurate reply in the peer frequency assessment (stages 3 and 6) was defined as a reply where the estimate was within a range of ±0.3 from the true mean.

6 In principle, the iteration could have been continued until consistency was achieved between subjective probability estimates and the frequency assessments, but we considered it likely that this would result in strategic behavior. In fact, one test subject did anticipate the fifth stage and acted strategically by inflating nine confidence intervals and making one point estimate. Unfortunately, this commendable clairvoyance was not rewarded since

samples t-test to compare population means between groups and a paired samples t-test to compare population means within groups.[7] To test if the population mean differs from 9 (the expected number of correct answers in stage 1 if the instructions are followed and there is no overconfidence) we used a Goodness-of-Fit test with the null hypothesis that the number of correct answers in stages 1 and 4 followed a binomial distribution with probability 0.9 and 10 independent draws. All reported p-values are two-sided.

### 2.2.1 Frequency assessments

In the experiment we measure overconfidence with both the standard interval method and with frequency assessments. We refer to overconfidence with these two methods as interval overconfidence and frequency overconfidence. Our first hypothesis to be tested is that interval overconfidence is greater than frequency overconfidence.

Hypothesis 1: Interval overconfidence is greater than frequency overconfidence.

$$\left|\mu_2^* - \mu_1^*\right| < \left|9 - \mu_1^*\right|, \left|\mu_5^* - \mu_4^*\right| < \left|9 - \mu_4^*\right|,$$

$$\left|\mu_2 - \mu_1\right| < \left|9 - \mu_1\right|, \left|\mu_5 - \mu_4\right| < \left|9 - \mu_4\right|.$$

The basis for this hypothesis is that we believe that participants knowingly, to some extent at least, set excessively narrow confidence intervals at the higher end of the response scale since they believe the social situation requires them to. When asked a question, people

---

the student was in the control group. Observations from this participant's answers in stages four and five were dropped from the sample.

typically assume that their primary knowledge is being scrutinized. Therefore, they will interpret the test as a test of primary knowledge and make a trade-off between informativeness and accuracy in trying to meet the standards of communication that they infer from the social situation. At the ninety percent level, this will manifest itself in a narrower interval than asked for. [8]

### 2.2.2 Peer frequency assessments

Drawing on the justification above, we also expect subjects to anticipate the interval overconfidence of their peers, since they are aware of the fact that they are partaking in a richer social setting where absolute adherence to explicit instructions is not the social norm. Thus, our second hypothesis is that the peer frequency assessments on average are below 9.

Hypothesis 2: Subjects anticipate the overconfidence of others.

$\mu_n^* < 9, \mu_n < 9$ for n=3, 6

### 2.2.3 Iteration

If subjects gave an inconsistent answer to the frequency question they were asked to repeat the interval estimation task. Iteration will have the effect of ingraining the notion that a ninety

---

[7] We also used non-parametric tests: the Wilcoxon test for paired samples and the Mann-Whitney test for independent samples. This led to similar results as the t-tests and does not change the reported conclusions below.

[8] Had our experiment asked for five percent confidence intervals, we would expect interval underconfidence.

percent interval really implies that nine questions should, on average, be answered correctly.[9] Thus, we anticipate that it will change the parameter values in the informativeness–accuracy trade-off in the intended direction (i.e. more consistent with the instructions). In addition, there is an element of learning, both because the instructions are reread and because the subjects are to some degree made aware of their failure to conform to the instructions. We therefore test the hypothesis that iteration will decrease interval overconfidence.

Hypothesis 3: Stationary replication without any feedback in the form of additional information will diminish interval overconfidence.

$$\left(9-\mu_1^*\right)>\left(9-\mu_4^*\right),\left(9-\mu_1\right)>\left(9-\mu_4\right).$$

### 2.2.4 Monetary incentives

Consistent with previous research (Davis and Holt, 1993; Smith, 1991; Smith and Walker, 1993) we expect monetary incentives to align stated judgments and genuine judgments. One simple reason is that subjects are likely to spend more cognitive resources on the task when good performance is financially rewarded. Another reason for providing economic incentives is that subjects may be unwilling to admit that they have not followed the instructions in the interval estimation task, and therefore exaggerate the frequency estimate to align it with the interval estimate. This tendency could be counteracted through monetary incentives, as self-validation is no longer costless. Thus, we test the hypothesis that monetary incentives will decrease frequency overconfidence.

[9] Even though this frequentist instruction is provided to the participants already in the initial instructions.

Hypothesis 4a: Monetary incentives will decrease frequency overconfidence.

$$\left(\mu_2 - \mu_1\right) > \left(\mu_2^* - \mu_1^*\right);\ \left(\mu_5 - \mu_4\right) > \left(\mu_5^* - \mu_4^*\right)$$

Previous research suggests that subjects may be more prone to admit the overconfidence of other subjects than their own overconfidence (Svenson, 1981). This is sometimes referred to as the 'above average effect'. Our final hypothesis to be tested is that the 'above average' effect will be smaller in the incentives group. The motivation for this hypothesis is that we expect monetary incentives to be more important for a subject's own frequency assessment than for the peer frequency assessment, as the hypothesized failure to comply with instructions will not bias the peer frequency assessment.

Hypothesis 4b: Monetary incentives will decrease the 'above average' effect.

$$\left(\mu_2 - \mu_3\right) > \left(\mu_2^* - \mu_3^*\right);\ \left(\mu_5 - \mu_6\right) > \left(\mu_5^* - \mu_6^*\right)$$

# 3. Results

Figure 1 illustrates the average results in the six stages of the experiment, for the control group and the incentives group. The average number of correct answers in the interval estimation task in stage 1 is 4.58 in the control group and 4.47 in the incentives group. The null hypothesis that the number of correct answers is 9 is clearly rejected ($p<0.001$). Having established that the test subjects' subjective confidence intervals were indeed incorrect (in

encompassing the correct value too seldom) we proceed by testing the data against our hypotheses.

## 3.1 Frequency assessments

As anticipated, the frequency assessments are inconsistent with the number of correct answers that the subjects were instructed to encompass with their intervals. On average subjects thought that the number of correct answers was 6.39 in the control group and 6.02 in the incentives group. The interval overconfidence is 4.42 in the control group and 4.53 in the incentives group. The frequency overconfidence is 1.82 in the control group and 1.55 in the incentives group. The difference in interval and frequency overconfidence is highly significant in both experimental groups ($p<0.001$), confirming hypothesis 1. Even though iteration decreases the interval overconfidence, the difference between the methods is still significant after iteration ($p=0.024$ in the control group and $p<0.001$ in the incentives group). It should also be noted that even though the use of frequency assessments decreases measured overconfidence, it does not eradicate the result entirely (overconfidence is significantly different from zero in both the incentives and the control groups in stages 2 and 5; $p<0.001$).

## 3.2 Peer frequency assessments

Given that subjects knowingly set too narrow confidence intervals, it is plausible that they will expect their peers to do the same. As seen in Figure 1 this is also the case. The average peer frequency assessment is 5.88 in the control group and 6.19 in the incentives group. This

differs significantly from 9 in both experimental groups ($p<0.001$), confirming hypothesis 2. The difference is significant also at stage 6 after iteration ($p<0.001$ in both groups).

## 3.3 Iteration

Our third hypothesis concerns iteration. The opportunity to revise the subjective confidence intervals in the iteration stage was exploited by 89 per cent of the participants in both groups. The average iteration effect (defined as the difference between the number of correct answers in stage four and stage one) was 1.37 in the control group and 1.15 in the incentives group. The interval overconfidence thus decreases with iteration consistent with hypothesis 3, and the effect is significant in both groups ($p<0.001$).

## 3.4 Monetary incentives

We hypothesized that monetary incentives would decrease frequency overconfidence. Before iteration monetary incentives decrease frequency overconfidence from 1.82 to 1.55, and after iteration monetary incentives decrease frequency overconfidence from 2.61 to 2.26. These differences are, however, not significant ($p=0.578$ before iteration and $p=0.426$ after iteration). We therefore cannot reject the null hypothesis of no effect of monetary incentives for hypothesis 4a.[10]

[10] One point concerning the interpretation of the data should be mentioned. If the perceived probability distributions in stages two and five are asymmetrically distributed around the modal observation, a wealth-maximizing individual will report the modal observation, whereas control group subjects may have been more inclined to provide their mean estimates.

We also hypothesized that monetary incentives would decrease the 'above average' effect. A visual inspection of Figure 1 does suggest that there is some merit to this view. Before iteration the 'above average' effect is 0.51 (p=0.027) in the control group and -0.17 (p=0.478) in the incentives group. This difference is significant, consistent with hypothesis 4b (p=0.040). After iteration the 'above average effect is 0.60 (p=0.005) in the control group and 0.27 (p=0.052) in the incentives group. This difference is smaller and not significant (p=0.188). Thus, we find some support for hypothesis 4b, but the results are not conclusive.

## 3.5 Gender differences

As part of the experiment we also collected information about the gender of subjects. We had no prior hypotheses about the directions of any potential gender differences. Figure 2 reports the results for the six stages of the experiment, disaggregated on gender. In the figure we have pooled the data for the control and incentives groups after testing for any interactions between gender and incentives (all the interaction effects were non significant, p>0.10). As can be seen in Figure 2 interval overconfidence in stages 1 and 4 are greater for women than for men, a difference which is weakly significant (p=0.095 before iteration and p=0.073 after iteration). Frequency overconfidence (the difference between stages 1 and 2 and between 4 and 5 in the Figure), however, is very similar for men and women, and does not differ significantly. This is illustrated more clearly in Figure 3, which shows interval and frequency overconfidence for men and women, respectively.

# 4. Discussion

DeBondt and Thaler (1995, quoted in Daniel et al., 1998, p. 1841) have argued that theories of psychological finance must be 'grounded on psychological evidence of how people actually behave'. The models of overconfidence developed in financial theory, are based on the assumption that agents not only respond, but also act overconfidently (Kyle and Wang, 1997; Odean, 1998). In justifying such an assumption, these authors habitually refer to psychological research, which investigates overconfidence in stated judgments. But it may not be appropriate to refer to research on judgment and take it as evidence of how people 'actually behave'. Our results show that people do not trust the accuracy of their own estimates of confidence intervals, nor the estimates of their peers. Using the frequency assessment method instead of the interval method reduced the measured overconfidence by about 60%. By providing monetary incentives in the frequency assessment method the overconfidence is reduced further (to about 35% of the original estimate), although this effect is not statistically significant in the experiment.

As measured overconfidence differs substantially between the interval method and the frequency estimation, a natural question is which response format elicits a more accurate measure of the subject's genuine subjective confidence (and can therefore be expected to have stronger behavioral consequences). Recent models of evolutionary psychology and laboratory work on the conjunction fallacy have both suggested that frequencies are the relevant input and response mode (Cosmides and Tooby, 1996; Gigerenzer et al., 1991; Hoffrage et al., 2000), and thus more interesting from a behavioral point of view. This is also supported by several results in the experiment. Firstly, subjects are aware of that they set too narrow confidence intervals in the interval method. Secondly, they are aware of that confidence

intervals provided by their peers cannot be interpreted as confidence intervals. Thirdly, when given an opportunity to revise the confidence intervals, most subjects widen the intervals. That we performed the test on subjects tutored in statistics is also advantageous since our results cannot be easily discarded on the grounds that our subjects were unfamiliar with the concept of a confidence interval. Our interpretation of why subjects knowingly provide too narrow confidence intervals is that they make a socially rational trade-off between accuracy and informativeness. This interpretation is consistent with the framework of conversational reasoning developed by Grice (1975) and the experimental evidence of Yaniv and Foster (1995, 1997). The notion that strict adherence to rules is not socially rational goes back at least to the political philosophy of Aristotle, who argued that effective practice cannot be reduced to pure rule-following.[11]

Authors who have carried out empirical research in interval estimation often highlight that the questions selected are taken from such domains where the participants 'ought' to know, according to some criterion. Thus, Russo and Schoemaker (1992, p. 9) emphasize the 'job relevance of the questions' used in their tests, and Sniezek and Buckley (1991) state that twelve financial officers in their study answered questions 'concerning the business operations of the university' (Sniezek and Buckley, 1991, p. 266). Such ambitions may exacerbate the misunderstanding of the task at hand, making subjects feel insecure. We do not contend that it is uninteresting to test the preciseness of knowledge of experts. However, it is an inevitable problem that knowledge questions given to professionals in their areas of expertise are likely to be interpreted as tests of primary knowledge, and thus ability. Previous research has indicated that experts are, in most cases, equally prone to biased judgments (Fischhoff and MacGregor, 1982). But such an assertion overlooks the fact that the job of experts is typically to forecast the value of a realized stochastic variable (e.g. What will

[11] See in particular books 2-4 of the Nicomachean Ethics in Aristotle; Crisp, (ed), 2000.

happen to the stock market tomorrow?) rather than the variance around the predicted outcome (e.g. Provide a 90% confidence interval for tomorrow's share price movements!). Consequently, it may be perfectly socially rational for an expert to knowingly provide an excessively narrow interval, given that it is reasonable to anticipate that strict adherence to the instructions will signal uncertainty and stupidity.

If interval overconfidence is partially attributable to a misapprehension of the task, then it follows that iteration should improve performance if it serves to ingrain the notion that 'an *x* percent confidence interval implies that in *x* percent of the cases, the true value should be within the stated range' and participants subsequently change their perception of the social situation. This learning process could be seen as a way of undoing a social desirability effect, in addition to allowing the subjects to think more carefully about the questions. Our results show that the subjects can significantly improve their assessment by a simple iteration in a test environment where they have very little to gain. It therefore seems likely that they should be able to improve the skill of providing correct intervals if they have sufficient incentive to do so in real life. It is perhaps no coincidence that the two groups that have demonstrated excellent calibration, meteorologists and horse track betters, both rely on this skill in their activities **(**Dowie, 1976; Murphy and Wrinkler, 1984**)**. It is possible that the correct interpretation of the results is not that managers lack metaknowledge and horse-betters do not, but rather that managers are less accustomed to expressing their knowledge in probabilistic terms. They have never needed to learn this simple skill.

It is important to relate our results to the existing work on overconfidence and calibration. Two papers are highly relevant for our study. Klayman et al. (1999) provide some comparative evidence on overconfidence in two-choice questions and subjective confidence tasks and find "little overconfidence" with two-choice questions and "pronounced

overconfidence" with confidence interval estimation.[12] Klayman et al. interpret these results as evidence of confidence judgments being "multiply determined" (Klayman et al., 1999, p. 217) and explain the huge overconfidence in interval estimation tasks by reference to 'biased retrieval and interpretation of evidence' (Klayman et al., 1999, p. 242). However, that explanation assumes that subjects are unaware of the allegedly biased nature of their cognitive processes, which is difficult to reconcile with the fact that subjects anticipated the interval overconfidence of others in our experiment. Sniezek and Buckley (1991) is the only previous study that compared the interval method with frequency assessments. In a study of twelve financial officers they found that the frequency assessments were highly inconsistent with the interval estimates, i.e. the frequency assessments produced lower overconfidence consistent with the results of our study. Sniezek and Buckley (1991, p. 263) interpreted this as reflecting 'unique psychological processes' in the different tasks. This is contrary to our explanation that subjects knowingly set too narrow intervals in the interval method.

Three main explanations have been prominent in the psychology literature to explain overconfidence: the heuristics and bias view (Kahneman et al., 1982), the statistical artefact view (Erev et al., 1994) and the ecological view (Gigerenzer et al.,1991; Juslin, 1993). The heuristics and biases view, pioneered by Kahneman and Tversky, attributes the overconfidence phenomenon to cognitive biases in human judgment (Kahneman et al., 1982). According to these authors, humans employ heuristics – cognitive tools that provide 'satisficing' solutions to complex problems – to reduce sophisticated tasks to 'simpler judgmental operations' (Kahneman et al., 1982, p. 3). Whereas a number of heuristics have been identified as important in explaining the general overconfidence result, one in particular is relevant in confidence interval estimation tasks, the so-called 'anchoring and adjustment heuristic'. However, anchoring cannot explain the difference observed in our experiment

[12] A two-choice question is where a subject has to choose one out of two mutually exclusive and universally

between the interval method and the frequency assessments, unless individuals are aware of that their interval estimates are excessively narrow. Similarly, the ecological and statistical artefact frameworks cannot easily account for this discrepancy.

The effect of providing monetary incentives was modest. The monetary incentives decreased frequency overconfidence somewhat, but not significantly so. The monetary incentives significantly weakened the 'above-average' effect in the first round, suggesting that this may be an artifact of using hypothetical questions.

We also investigated gender differences in overconfidence. Interestingly the interval method indicated that women are significantly more overconfident than men, whereas the frequency estimation indicated a similar level of overconfidence for both men and women. One interpretation of this result is that the parameterization of the trade-off between accuracy and informativeness differ between men and women, rather than the level of overconfidence. This finding may be of relevance for the interpretation of cross-cultural differences in overconfidence in previous studies. A number of studies have investigated cross-cultural differences in overconfidence using the interval method, most often by comparing Western students with Asian students (see Yates et al. 1989; Yates et al. 1996). The typical finding has been that Asian students exhibit greater overconfidence, which has been taken as evidence that Asians are more overconfident. From our perspective, it seems at least equally reasonable to expect the observed differences to reflect variations in the trade-off between accuracy and informativeness, rather than genuine differences in overconfidence. Further work using frequency assessments comparing subjects across cultures is needed to settle this issue.

---

exhaustive answers and then has to state her confidence in the chosen answer (usually in percent).

# 5. Conclusion

Edwards and von Winterfeldt (1986, p. 127) once remarked that 'Research evidence about calibration is abundant but singularly hard to make sense of'. It was in this spirit that we set out to investigate the discomforting result that people fail miserably in confidence interval estimation tasks. We demonstrate that a large share of the calculated overconfidence in interval estimation tasks is attributable to the difference between subjective probability estimates and frequency assessments. Our results suggest that the interval estimation method suffers from several shortcomings. Subjects do not, we argue, appreciate the metacognitive nature of the interval estimation task, and in seeking to make an informative contribution therefore narrow their intervals excessively so as to not signal ignorance. Consequently, it is the peculiarity of the response format, and not a genuine cognitive bias, that is the chief culprit. The suggestion that frequencies are the relevant response format is also consistent with insights from evolutionary and experimental psychology.

In conclusion therefore, French taxi-drivers and subjects in interval estimation tasks employ similar strategies and unless the subjects should decide to go on strike – something we consider rather unlikely – confidence interval estimation tasks will remain flawed as a measure of overconfidence. Further work is needed to establish the behavioral consequences of overconfidence before it can be recognized as an important departure from homo oeconomicus.

# REFERENCES

Alpert, M. and Raiffa, H. (1969). A Progress Report on the Training of Probability Assessors. Unpublished manuscript. Reprinted in: Kahneman, D., Slovic, P. and Tversky, A. (eds.) (1982). Judgment Under Uncertainty: Heuristics and Biases. Cambridge: Cambridge University Press, pp. 294-305.

Babad, E. (1987). Wishful Thinking and Objectivity Amongst Sports Fans. Social Behaviour 2, 231-240.

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

Camerer, C. F. and Hogarth, R. M. (1999). The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework. Journal of Risk and Uncertainty 19, 7-42.

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

Crisp, R. (ed.) (2000). Nicomachean Ethics. Cambridge: Cambridge University Press.

Daniel, K., Hirshleifer, D. and Subrahmanyam, A (1998). Investor Psychology and Security Market Under- and Overreactions. Journal of Finance 53, 1839-1885.

Davis, D. and Holt, C. (1993). Experimental Economics. Princeton: Princeton University Press.

DeBondt, W. F .M. and Thaler, R. H. (1995). Financial Decision-Making in Markets and Firms: A Behavioral Perspective. In: Jarrow, R. A., Maksimovic, V. and Ziemba, W. T. (eds.). Finance, Handbooks in Operations Research and Management Science, Vol. 9, Chapt. 13. Amsterdam: North Holland, pp. 385-410.

DeLong, J., Shleifer, A., Summers, L. and Waldmann R. (1991). The Survival of Noise Traders in Financial Markets. Journal of Business 64, 1-19.

Dowie, J. (1976). On the Efficiency and Equity of Betting Markets. Economica 43, 139-150.

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

Fischhoff, B. and MacGregor, D. (1982). Subjective Confidence in Forecasts. Journal of Forecasting 1, 155-172.

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

Grice, P. (1975). Logic and Conversation. In: Cole, P. and Morgan, J. (eds.). Syntax and Semantics, Vol. 3: Speech Acts. New York: Academic Press, pp. 41-58.

Hertwig, R. and Ortmann, A. (2001). Experimental Practices in Economics: A Methodological Challenge for Psychologists? Behavioral and Brain Sciences 24, 383-451.

Hertwig, R. and Gigerenzer, G. (1999). The 'Conjunction Fallacy' Revisited: How Intelligent Inferences Look Like Reasoning Errors. Journal of Behavioral Decision Making 12, 275-305.

Hoffrage, U., Lindsey, S., Hertwig, R. and Gigerenzer, G. (2000). Communicating Statistical Information. Science 290, 2261-2262.

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

Juslin, P. (1993). An Explanation of the Hard-Easy Effect in Studies of Realism of Confidence of One's General Knowledge. European Journal of Cognitive Psychology 5, 55-71.

Kahneman, D., Slovic, P., and Tversky, A. (eds.) (1982). Judgment under Uncertainty: Heuristics and Biases. Cambridge: Cambridge University Press.

Keren, G. (1991). Calibration and Probability Judgments: Conceptual and Methodological Issues. Acta Psychologica 77, 217-273.

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

Kyle, A. and Wang, F. A. (1997). Speculation Duopoly With Agreement to Disagree: Can Overconfidence Survive the Market Test? Journal of Finance 52, 2073-2090.

Lichtenstein, S., Fischhoff, B., and L. D. Phillips (1982). Calibration of Probabilities: The State of the Art to 1980. In: Kahneman, D., Slovic, P. and Tversky, A. (eds.). Judgment Under Uncertainty: Heuristics and Biases. Cambridge: Cambridge University Press, pp. 306-334.

May, R. S. (1986). Overconfidence as a Result of Incomplete and Wrong Knowledge. In: R. W. Scholz (ed.). Current Issues in West German Decision Research. New York: Lang, pp. 13-30.

Moskowitz, T. J. and Vissing-Jorgensen, A. (2002). The Returns to Entrepreneurial Investment: A Private Equity Premium Puzzle?. American Economic Review 92, 745-778.

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

Myers, D. G. (1993). Social Psychology, 4th edition. New York: McGraw-Hill.

Odean, T. (1999). Do Investors Trade Too Much? American Economic Review 89, 1279-1298.

Odean, T. (1998). Volume, Volatility, Price and Profit When All Traders Are Above Average. Journal of Finance 53, 1887-1934.

Orphanides, A. (2000). The Quest for Prosperity Without Inflation. European Central Bank Working Paper Series, 15.

Russo, J. E. and Schoemaker, P. J. H. (1992). Managing Overconfidence. Sloan Management Review 33, 7-17.

Scott, J. (1998). Seeing Like a State. New Haven: Yale University Press.

Smith, V. (1991). Rational Choice: The Contrast Between Economics and Psychology. Journal of Political Economy 99, 877-897.

Smith, V. and Walker, J. (1993). Monetary Rewards and Decision Cost in Experimental Economics. Economic Inquiry 31, 245-261.

Sniezek, J. and Buckley, T. (1991). Confidence Depends on the Level of Aggregation. Journal of Behavioral Decision Making 4, 263-272.

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

Tversky, A. and Kahneman, D. (1983). Extensional vs. Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. Psychological Review 90, 293-315.

Von Winterfeldt, D. and Edwards, W. (1986). Decision Analysis and Behavioral Research. Cambridge: Cambridge University Press.

Yaniv, I. and Foster, D. (1995). Graininess of Judgment Under Uncertainty: An Accuracy Informativeness Tradeoff. Journal of Experimental Psychology: General 124, 424-432.

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

Yates, J. F. (1990). Judgment and Decision Making. Englewood Cliffs, NJ: Prentice Hall.

Yates, J. F., Lee, J. W. and Shinotsuka, H. (1996). Beliefs About Overconfidence, Including Its Cross National Variation. Organizational Behavior and Human Decision Processes 65, 138-147.

Yates, J.F., Zhu, Y., Ronis, D., Wang, D., Shinotsuka, H. and Toda, M. (1989). Probability Judgment Accuracy: China, Japan and the United States. Organizational Behavior and Human Decision Processes 43, 145-171.

# FIGURES

*Figure 1. Average results in the six stages of the experiment, disaggregated on test conditions.*

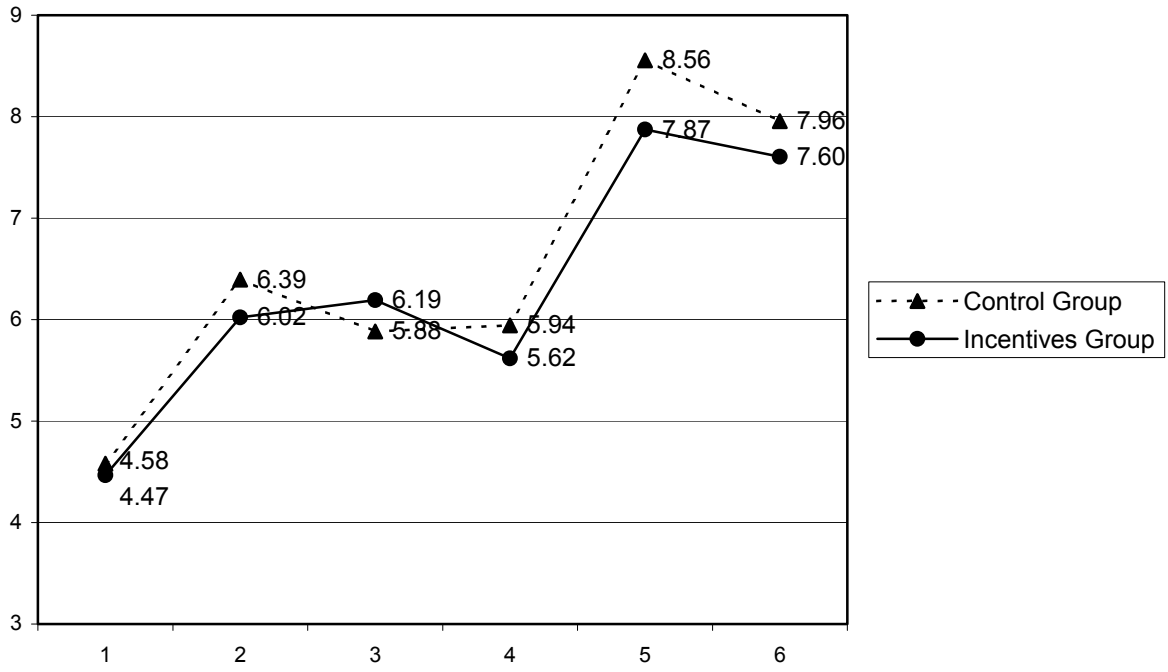


*Figure 2. Average results in the six stages of the experiment, disaggregated on gender.*

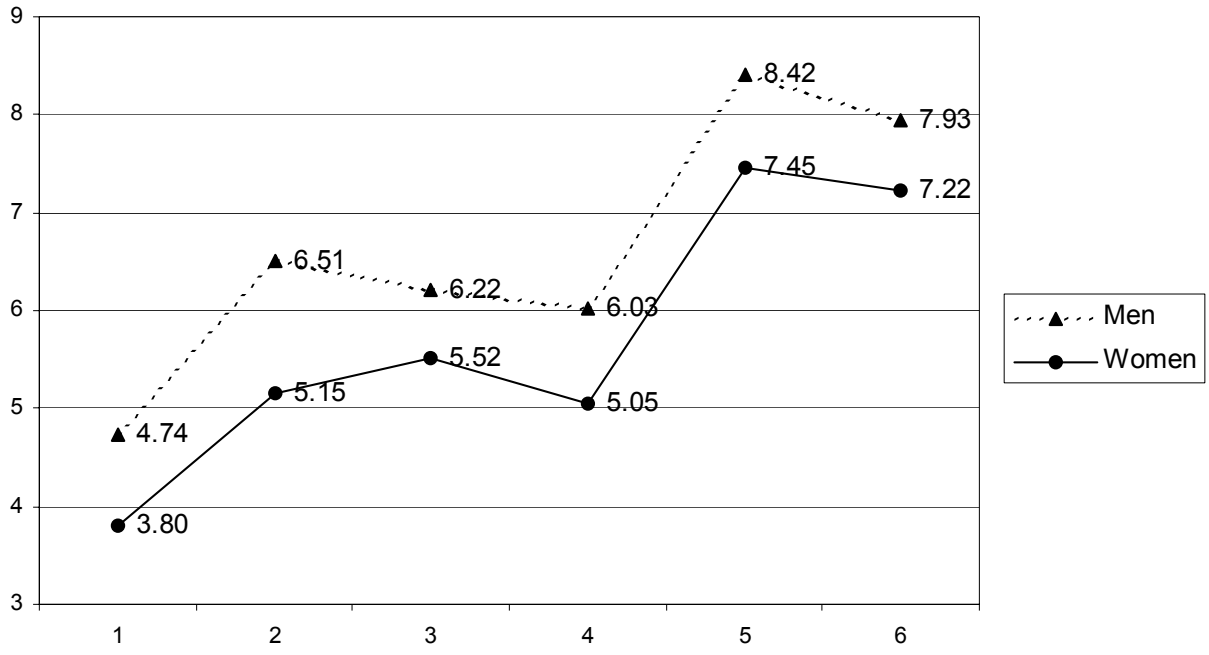
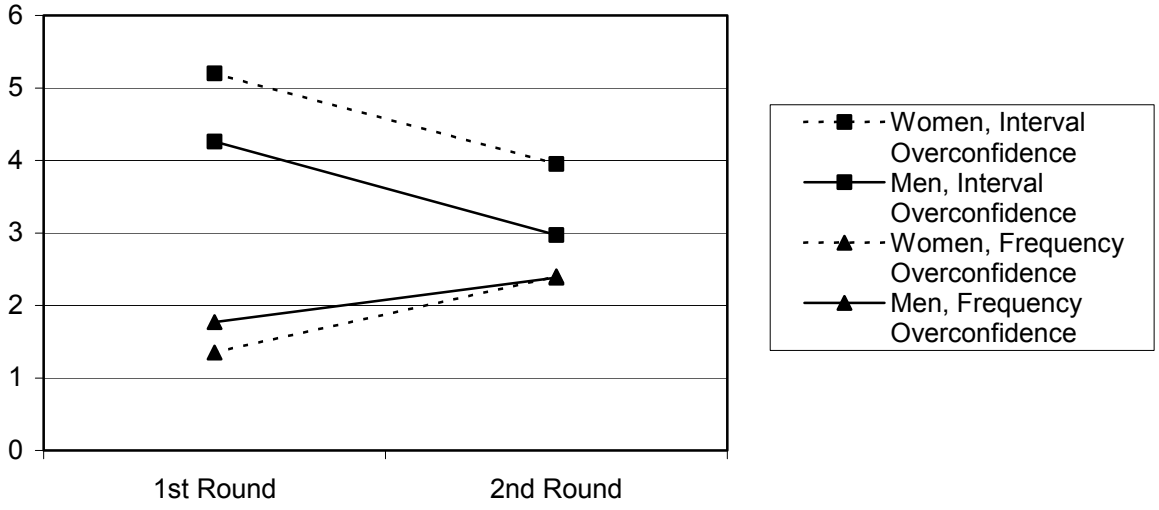


*Figure 3. Frequency and interval overconfidence in the first and second rounds, disaggregated on gender.*



## APPENDIX: EXPERIMENTAL INSTRUCTIONS

The original instructions were in Swedish. Text in brackets, [], is here used to denote words that varied between participants. Text below which is presented in bold was not included in the original text.

### PAGE 1

General instructions:

- This experiment will be conducted in a number of Stages, where each Stage corresponds to a sheet with questions.
- After completion of the experiment, we will correct your answers. You will also receive a solution key immediately upon finishing the experiment.
- After completing a Stage the answer is final, hence you may not change any answers on completed Stages.
- Write legibly!
- Kindly state your age…….and gender…….
- Your answers will be considered confidential and processed as such!

### PAGE 2

Question sheet [participant number]

1. What was the population of [Swedish municipality] in [year]?
2. What was the turnover of stocks traded on the A-list of Stockholm Stock Exchange in [year] (in billions of Swedish kronor)?
3. What was the development of the SAX index in [year] (in percent)?
4. What was the average monthly salary of a [gender] public employee, age [age group] with [education level] in the year 2001?
5. What was the value of Swedish exports (in billions of Swedish kronor) to [country] during the period January-October [2001 or 2002]
6. What was the price in US dollars of [commodity] in [month] [year]?
7. What were the pre-tax earnings (in millions of Swedish kronor) of [Swedish company] in [year]?
8. What was the size of the Swedish national debt (in billions of Swedish kronor) in [month] [year]?
9. What was the price of a US dollar in Swedish kronor in [month] [year]?
10. How many seats did the [party] capture in the **(Swedish)** parliamentary elections of [year]?

### PAGE 3

Stage 1

Your task is to provide 90 percent confidence intervals for the answers to the ten questions found on the sheet with the title ‘Question sheet’. A confidence interval is an interval which

contains a certain unknown value with a certain probability. In other words, for every question you should provide an upper and a lower limit, such that you assess the probability to be 90 percent that the correct answer is between the two limits. Hence, you should aim at answering correctly in 90 percent of the cases and incorrectly in 10 percent of the cases.

The questions are randomly chosen from a pool of possible questions and are thus not intended to be "tricky" or misleading. (For example, this means that in question 1, the computer randomly chose the municipality and the year that we ask for.) PAY CLOSE ATTENTION TO THE UNITS! For example, distinguish between millions and billions!

Example question:

*What was the population of Great Britain in 1997 (in millions)?*

If you are 90 percent certain that the population of Great Britain in 1997 was between 47 000 000 and 80 000 000 , you write:

| Lower limit | Upper limit |
|---|---|
| 47 | 80 |

[Table with 20 blank cells for the lower and upper limits of the ten questions]

When you are ready, place your pen on the desk in front of you.

## PAGE 4 (for the incentives group)

Stage 2

Look back at your answers from Stage 1. Without changing any of the stated intervals, estimate how many of these intervals you believe contain the true value. In other words, how many correct answers do you think you had in Stage 1?

Answer: …….correct answers

A correct answer in this Stage will be rewarded with 50 Swedish kronor **(approximately 6 US dollars).** The stated intervals in Stage 1 do not affect the reward, the only criterion is whether or not you make an accurate assessment of the number of correct answers in Stage 1. It does not matter if you have many or few correct answers, as long as your estimate is consistent with the actual number of correct answers.

(Participants who answer correctly in this Stage will be notified by e-mail this afternoon and can collect their reward in the school pub tomorrow between 1 pm and 1.15 pm. If you cannot attend at that time you can answer by e-mail and we will arrange an additional opportunity to collect the reward.)

## PAGE 4 (for the control group)

Stage 2

Look back at your answers from Stage 1. Without changing any of the stated intervals, estimate how many of these intervals you believe contain the true value. In other words, how many correct answers do you think you had in Stage 1?

Answer: …….correct answers

It does not matter whether you have many or few correct answers, all that matters is that your estimate agree with the true number of correct answers.

When you are ready, place your pen on the desk in front of you.

## PAGE 5 (for the incentives group)

Stage 3

All participants in the experiment receive the same instructions as you do. The questions in Stage 1 are not identical for all participants, but they are similar and of comparable difficulty. The questions are generated randomly.

State with one decimal the average number of correct answers that you think the other participants in the experiment have managed to capture with their confidence intervals in Stage 1. Here, too, a correct answer will be rewarded with 50 Swedish kronor **(approximately 6 US dollars).** An answer will be judged correct if it is within ±0.3 of the exact mean. The only condition for being granted the reward is whether your estimate below is consistent with the true average.

Answer: …….correct answers

(Here, too, the participants who gave a correct answer can collect their reward in the school pub between 1 pm and 1.15 pm or make a separate arrangement by e-mail.)

When you are ready, place your pen on the desk in front of you. After that, place your answers from Stages 2 to 3 in the envelope in front of you and await further instructions.

## PAGE 5 (for the control group)

Stage 3

All participants in the experiment receive the same instructions as you do. The questions in Stage 1 are not identical for all participants, but they are similar and of comparable difficulty. The questions are generated randomly.

State with one decimal the average number of correct answers that you think the other participants in the experiment have managed to capture with their confidence intervals in Stage 1.

Answer: …….correct answers

When you are ready, place your pen on the desk in front of you. After that, place your answers from Stages 2 to 3 in the envelope in front of you and await further instructions.

## PAGE 6

Stage 4

Adjust your confidence intervals so that you really believe they capture the correct answer in 90 % of the cases. You may look back at your answers from Stage 1 for guidance. We would however like to remind you that you are not allowed to change any of your earlier answers. Write the new confidence intervals below:

[table with 20 blank cells for the lower and upper limits of the ten questions]

When you are ready, place your pen on the desk in front of you.

## PAGE 7 (for the incentives group)

Stage 5

Look back at your answers from Stage 4. Without changing any of the stated intervals, estimate how many of these intervals you believe contain the true value. In other words, how many correct answers do you think you had in Stage 4?

Answer: …….correct answers

A correct answer in this Stage will be rewarded with 50 Swedish kronor **(approximately 6 US dollars).** The stated intervals in Stage 4 do not affect the reward, the only criterion is whether or not you make an accurate assessment of the number of correct answers in Stage 4. It does not matter if you have many or few correct answers, as long as your estimate is consistent with the actual number of correct answers.

(Here, too, the participants who gave a correct answer can collect their reward in the school pub between 1 pm and 1.15 pm or make a separate arrangement by e-mail.)

## PAGE 7 (for the control group)

Stage 5

Look back at your answers from Stage 4. Without changing any of the stated intervals, estimate how many of these intervals you believe contain the true value. In other words, how many correct answers do you think you had in Stage 4?

Answer: …….correct answers

It does not matter whether you have many or few correct answers, all that matters is that your estimate agree with the true number of correct answers.

When you are ready, place your pen on the desk in front of you.

## PAGE 8 (for the incentives group)

Stage 6

State with one decimal the average number of correct answers that you think the other participants in the experiment have managed to capture with their confidence intervals in Stage 4. Here, too, a correct answer will be rewarded with 50 Swedish kronor **(approximately 6 US dollars).** An answer will be judged correct if it is within ±0.3 of the exact mean. The only condition for being granted the reward is whether your estimate below is consistent with the true average.

Answer: …….correct answers

(Here, too, the participants who gave a correct answer can collect their reward in the school pub between 1 pm and 1.15 pm or make a separate arrangement by e-mail.)

When you are ready, place your pen on the desk in front of you. Place all sheets in the envelope in front of you. Write your student number on the envelope. (So that we can contact you regarding possible additional remuneration.)

## PAGE 8 (for the control group)

Stage 6

State with one decimal the average number of correct answers that you think the other participants in the experiment have managed to capture with their confidence intervals in Stage 4.

Answer: …….correct answers

When you are ready, place your pen on the desk in front of you. Place all sheets in the envelope in front of you.