University of Warsaw
Faculty of Economic Sciences

Working Papers

No. 15/2011 (55)

Michał Krawczyk

Overconfident for real? Proper scoring for confidence intervals

Warsaw 2011

University of Warsaw
Faculty of Economic Sciences

Overconfident for real?
Proper scoring for confidence intervals

Michał Krawczyk
University of Warsaw
Faculty of Economic Sciences
e-mail: mkrawczyk@wne.uw.edu.pl

Abstract

Studies show that people tend to provide overly narrow confidence intervals for unknown values. Such a form of overconfidence would have an important impact on financial markets, among other domains, leading i.a. to excessive trading. The present study is one of the very few that try to incentivize reporting correct confidence intervals. To this end, a reward scheme is proposed, based on a combination of asymmetric loss functions minimized by appropriate quantiles of a probability distribution. In the experiment I find that incentivized subjects provide wider confidence intervals, obtaining a higher hit rate than the control group. The effect is stronger than that of feedback and explicit warning. These findings suggest that the overly narrow confidence intervals reported elsewhere are partly due to an insufficient mental effort that subjects exert and that they can be induced to do so by the proposed incentive scheme.

Keywords:
overconfidence, calibration, confidence intervals, proper scoring rules

JEL:
C44, C91, D03, D84, G17

1 Introduction

Everyday conversations often involve uncertain values expressed in terms of intervals (such as "a women in her thirties was here between 4:00 and 4:30, looking for you") that are expected to cover the true value with high probability. Also experts in various domains are frequently asked to make range predictions or judgments, perhaps at pre-specified levels of confidence. The method is useful because it gives a sense of degree of uncertainty involved, without forcing those generating or using the prediction to deal with entire subjective probability distributions. Experimental research in psychology shows, however, that people tend to be overconfident about the precision of their knowledge and ability to predict. In a typical design, subjects would be asked to submit a 90% confidence interval (CI) for a value that they are believed not to know precisely (such as the length of the Rhine) or a value that cannot be known yet (such as the value of Dow Jones a month later). It is usually found that the intervals are too narrow on average as few as about 50% may actually cover the relevant value. This form of overconfidence appears to be particularly robust (Klayman, Soll, González-Vallejo, and Barlas 1999).[1] In particular, studies tend to conclude that experts (McKenzie, Liersch, and Yaniv 2008), particularly professional traders (Oberlechner and Osler 2008) and CFOs (Ben-David, Graham, and Harvey 2010) do as poorly as students or worse. Consequently, interval overconfidence has been proposed to have an important impact on financial markets. Theoretical models allowing underestimation of information signal variance, which is parallel to overly narrow CIs, can, among other things, explain excessive trading (see e.g. Glaser, Langer, and Weber (2007) for an overview). Empirically, Biais, Hilton, Mazurier, and Pouget (2005) find that overconfident subjects perform worse in an experimental market. On the other hand, the hypothesized link between interval overconfidence and trade volume at the individual level was not present in the field study of professional traders (Glaser and Weber 2007).[2] Recently, the impact of biased interval judgment on corporate finance is beginning to be investigated. Ben-David, Graham, and Harvey (2010) find that firms with overconfident financial executives tend to invest more.

There are numerous studies aimed at reducing overconfidence in intervals by manipulating the details of the elicitation mode (e.g. asking subjects to additionally provide their best guess, see Speirs-Bridge, Fidler, McBride, Flander, Cumming, and Burgman (2010), Soll and Klayman (2004) and re-

[1] See, however, (Budescu and Du 2007) that only finds slight overconfidence at 90% confidence level and slight underconfidence at .50% level.

[2] In contrast, Grinblatt and Keloharju (2009) find that a generic assessment of overconfidence is predictive of subsequent trading volume in a large sample of Finnish investors.

quiring probability to be assigned to each possible interval, Haran, Moore, and Morewedge (2010) or by asking to think about "improbably" high and low values, see (Teigen and Jørgensen 2005), experiment 4) or by providing detailed feedback (Bolger and Onkal-Atay 2004). While some beneficial effects of these methods are well-documented, they do not seem to reduce the bias entirely; they generally lack formal theoretical underpinning; their relevance to some real-world applications may be limited; finally, there is always a threat that they may lead to "overshooting" – as long as the sample is strongly overconfident, *any* manipulation leading to wider intervals will "improve calibration" as measured at group level (see Murphy and Stevens (2004), for a similar debate on hypothetical bias).

While many different remedies have been proposed, it remains unknown whether or not sufficient *monetary incentives* could reduce the bias – nearly all of controlled studies on interval overconfidence involved hypothetical questions.[3] Overconfidence outside of the lab, on the other hand, may be quite costly to the decision maker.

The goal of this study is to establish whether the tendency to set overly narrow CIs indeed persists when an individual faces financial incentives to report them carefully. Clearly, providing CIs reflecting one's best judgment requires certain mental effort that individuals may not be willing to make. Furthermore, one can be more proud getting it right and less ashamed missing it after having submitted quite a narrow interval.[4] The general trade-off between accuracy and informativeness inherent in interval estimation tasks has been studied i.a. by Yaniv and Foster (1995). Until now, however, to the best of my knowledge, no simple strategy to incentivize CIs has been tested. [5] This is in stark contrast to a very closely related task, namely probability judgments, where proper scoring rules (PSR) have been known for several

[3]Most authors confine themselves to the remark that direct rewarding of proper calibration (e.g. having nine out of 10 intervals actually cover the true values when the required confidence level is 90%) should lead sufficiently smart subjects to provide nine unreasonably wide intervals and one very narrow. A much more subtle dynamic method of interval width adjustment may be applied when judges receive immediate feedback; it may well be used by experts in the field, trying to appear properly calibrated.

[4]A software developer told Jørgensen, Teigen, and Moløkken (2004) "I feel that if I estimate very wide [confidence intervals for work hours required in the project], this will be interpreted as a total lack of competence [...] I'd rather have fewer actual values inside the minimum-maximum interval, than providing meaningless, wide [intervals]." See also Study D in the same paper and extreme miscalibration of software developers reported in (Connolly and Dean 1997). Polish Central Bank, on the other hand, seems to be immune to this kind of concerns. In February 2010 it predicted at just 50% confidence level (though correctly, as it later turned out) that the country's 2010 GDP growth rate would be between 2.1 and 4.1.

[5]See however (Schlag and Weele 2009) and studies cited therein for a related approach.

decades (Brier 1950) and used in numerous studies. These tend to show that the use of PSR improves performance, in a sense of avoiding overly low probabilities being assigned to particular outcomes (which may be heavily penalized by PSR). More generally, reviews such as (Smith and Walker 1993) prove that incentives often make a difference in economic experiments.

The study that is perhaps closest to mine is (Van Lenthe 1993) that actually uses PSR for continuous distributions to encourage faithful reporting of CIs. To this end, the author asks the subjects for the lower bound, best guess and higher bound for a number of variables expressed as fractions. Then he assumes that the underlying distribution is a beta distribution, estimates it given the three reported values and scores using the PSR. Obviously, the assumption of beta distribution is problematic, and it is not clear how this kind of approach should be generalized to variables with support beyond the unit interval. [6] Another study that tried to incentivize proper calibration for intervals was (Dargnies and Hollard 2008). The authors find the impact of an incentivized calibration training session in men only but even then it does not seem to be significant. The reward scheme is not incentive compatible and, curiously enough, the control group subjects were told "they would receive remuneration regarding this task but that they would only know how the remuneration was established later."

In this study, I develop and implement a simple scoring rule for CIs that is relatively easy to comprehend and puts no assumptions on subjects' subjective probability distributions.

The experiment confirms the efficacy of the proposed incentivization procedure - it makes subjects' reported CIs 26% wider. As a result overconfidence (the difference between the required confidence level of 90% and the frequency with which the reported intervals actually cover the unknown real value, henceforth: the hit rate) is reduced by 14 pct. points. These treatment effects are comparable or stronger than the impact of other investigated interventions (namely explicit warnings and feedback) and continue to operate when combined with them. However, even when all three measures are applied, subjects remain overconfident and there is no evidence of any of them reporting overly wide intervals.

2 Proper scoring rule for confidence intervals

Consider an individual holding a subjective belief about an unknown value described by a cumulative probability distribution function $F(X)$. The indi-

[6] Additionally, the standard approach in the overconfidence research tradition is to ask for the bounds only, not the most likely case.

vidual is asked to report the $\alpha/2$ and the $1-\alpha/2$ quantiles of the distribution, denoted Q_L and Q_H respectively,[7] for some $\alpha \in (0,1)$. Next, a single realization of the variable following the subjective distribution (i.e. the true value) is observed. The goal is to provide a loss function that will encourage truthful reporting. Thus, if the reported values are x_L and x_H respectively and the observed realization is x_0, how should the payment for the individual, $L(x_L, x_H, x_0)$ be determined in order to induce reporting $x_L = Q_L$ and $x_H = Q_H$?

Formally, it should be the case that

$$\begin{aligned} argmax_{x_H} \int_{x_0=-\infty}^{x_0=+\infty} L(x_L, x_H, x_0) dF(x_0) &= Q_H \\ argmax_{x_L} \int_{x_0=-\infty}^{x_0=+\infty} L(x_L, x_H, x_0) dF(x_0) &= Q_L \end{aligned} \tag{1}$$

Fact 1 *For a risk-neutral subject, the loss function,* $L(x_L, x_H, x_0) = -\frac{\alpha}{2}(x_H - x_L) - 1_{(x_0 > x_H)}(x_0 - x_H) - 1_{x_0 < x_L}(x_L - x_0)$ *induces truthful reporting of the symmetric* $1-\alpha$ *confidence interval.*[8]

Proof Given subjective beliefs and the loss function, reporting any x_L, x_H leads to an expected payoff of

$$E(\Pi) = \int_{x_0=-\infty}^{x_0=+\infty} (-\frac{\alpha}{2}(x_H - x_L) - 1_{(x_0 > x_H)}(x_0 - x_H) - 1_{x_0 < x_L}(x_L - x_0)) dF(x_0) \tag{2}$$

Taking the first derivative with respect to x_L, x_H yields the First Order Conditions

$$\begin{aligned} \frac{\partial E(\Pi)}{\partial x_L} = \int_{x_0=-\infty}^{x_0=+\infty} (\frac{\alpha}{2} - 1_{x_0 < x_L}) dF(x_0) &= 0 \\ \frac{\partial E(\Pi)}{\partial x_H} = \int_{x_0=-\infty}^{x_0=+\infty} (-\frac{\alpha}{2} + 1_{x_H < x_0}) dF(x_0) &= 0 \end{aligned} \tag{3}$$

Thus:

$$\begin{aligned} \frac{\alpha}{2} &= F(x_L) \\ \frac{\alpha}{2} &= 1 - F(x_H). \end{aligned} \tag{4}$$

[7] In the case of a non-convex support resulting in an interval rather than single point satisfying $F(X) = \alpha/2$, we take Q_L to be the the sup of this set. Similarly Q_H is the inf over the set of values for which $F(X) = 1 - \alpha/2$.

[8] Jose and Winkler (2009) make the same observation, additionally providing a more general formula, applicable to different sets of quantiles.

as required and the SOC is obviously satisfied. QED.

The proposed function is obviously not unique but it can be shown that it is unique up to trivial transformations in the class of functions additively separable in x_L and x_H. While fairly simple (piece-wise linear in each argument and, for any (x_L, x_H) interval, symmetric around its midpoint), the scheme may seem unnatural at first. However, it has close analogues outside of the laboratory. For example, it is basically equivalent to the choice of strike prices in the short strangle option strategy, in which the investor writes out-of-the-money put and call options for the same maturity and underlying asset, see Figure 1 This strategy is attractive if the investor expects relatively little volatility, i.e. is quite confident about the future price of the asset. The greater the difference between the strike prices, the lower the option premiums she will earn (the only substantial difference wrt the proposed reward scheme being that this relationship is not linear). Her losses in the case of large price movements are generally unlimited.

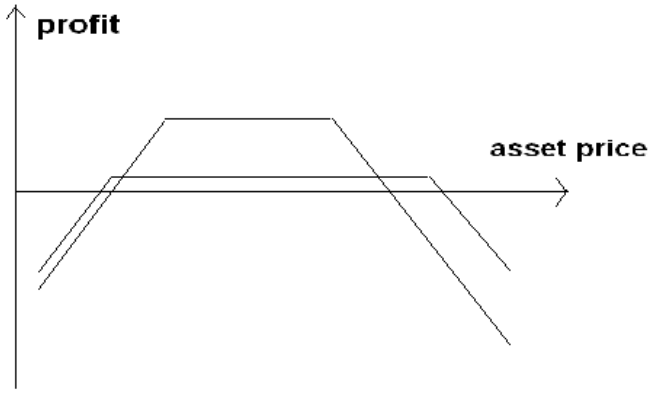

Figure 1: Comparison of two short strangle strategies

Examples of similar loss functions penalizing careless setting of quantiles of one's subjective distribution also abound in other life domains. A typical cell phone tariff includes a flat fee and relatively steep per-minute charges. The higher the fee, the greater the number of "included minutes". Grubb (2009) is able to explain ex post tariff choice "mistakes" in this market in terms of customer overconfidence. By the same token, when driving to an important meeting we implicitly trade off the likely minor inconvenience of

having to wait for the others against the potentially very costly possibility of coming too late. We are thus trying to leave enough time for the latter to be quite unlikely (but many of us fail, i.e. find themselves being late much more often than they would like to, another case for overconfidence).[9] A farmer may underinsure against drought and flood if he overestimates precision of his precipitation forecasts. A manager responsible for the supply chain may fail to provide sufficient warehouse capacity and emergency supply sources given demand uncertainty and possible delays of the primary supplier etc.

The properness of the proposed scoring procedure rests on the assumption of risk neutrality. Several studies show that some people violate it, even for small stakes, and may behave in ways inconsistent with the expected utility theory.

It can be shown that for any given subjective beliefs about the actual value and any two intervals of which one is a proper subset of the other, the former results in a more risky gamble, in terms of second-order stochastic dominance of deviations from the respective means. If, therefore, subjects are risk-averse, we expect them to provide wider intervals that will cover the actual value more often than 90% of the time. Prospect theory of Kahneman and Tversky (1979) proposes that people will tend to overweight the probability of unlikely events yielding particularly high or particularly low outcomes. In our case, it will mean paying additional attention to the possibility of not covering the actual value, which will again result in wider intervals.

The proposed procedure may be generalized to account for such deviations from risk neutrality in a way similar to (Offerman, Sonnemans, Van de Kuilen, and Wakker 2009). For this purpose, subject-level calibration could be performed involving elicitation of $1-\alpha$ confidence intervals for a variable known to follow a uniform distribution, say, on $[0,1]$. Such $\alpha_L/2$ and $\alpha_H/2$ need to be found that subject would report the correct interval $[0.05, 0.95]$ when confronted with the loss function $L(x_L, x_H, x_0) = -\frac{\alpha_H}{2}x_H + \frac{\alpha_L}{2}x_L - 1_{(x_0>x_H)}(x_0 - x_H) - 1_{x_0<x_L}(x_L - x_0)$ (which could possibly be rearranged to make it more transparent to the subjects). We would then know that the L function should induce reporting correct $1-\alpha$ intervals for variables following unknown subjective distribution. The question whether this additional layer of complexity is rewarded with significantly better calibration is an empirical one and is left for future research.

[9] In some of such examples, the difference wrt. to the discussed reward scheme is that we focus on one quantile only e.g. the driving time that is unlikely to be *exceeded*; additionally, there may be non-linearity involved, e.g. it may not matter much any more whether someone is 30 or 40 minutes late for a job interview.

3 Experimental design

The experiment seeking to verify the usefulness of the incentivizing scheme described above was implemented as follows. The experiment was advertised among registered subjects of the Laboratory of Experimental Economics at the University of Warsaw using ORSEE (Greiner 2004). Upon arriving in the lab, each subject was randomly assigned to one of two treatments: the Proper Scoring Treatment (PST) or Control Treatment (CT). Subjects were not aware of the existence of the non-assigned treatment. All subjects read the instructions explaining the task of providing 90% CIs for distances in kilometers between twenty pairs of major European cities.[10][11] The English translation of the instructions is provided in the appendix. Subjects in the PST additionally learned about the loss function described in the previous section. The losses were expressed in points, whereby 1000 points would corresponded to 4 PLN (approx. 1 euro). For example, giving an interval of (1000,1400) for the distance between Barcelona and Stockholm would result in a loss of 20 points due to the interval width ($.05 \times 400$) and 880 points due to missed true value (2280 kilometers), to make an aggregate loss of 900 points or 3.6 PLN. Subjects in the PST would start the experiment with 12'000 points[12] with no additional show-up fee, while subjects in the control treatment would earn 30 PLN for sure.

Having read the instructions, subjects would start their z-tree (Fischbacher 2007) treatment, beginning with three trial periods. Next, three control questions were asked, whereby subjects could only proceed by providing the correct answer. Subjects were finally asked whether they were sure they understood the task and could then go on with providing 20 CIs for different distances. For the first 10 questions, no feedback was given. After this block, subjects were asked to guess how many of their 10 intervals covered the actual values. This question was not incentivized, for it would interfere with behavior the interval setting task itself ("moral hazard problem" which may

[10] All subjects within a session were given the same questions in the same order, to obtain maximum power of the treatment comparison. For each new session a new set of questions was randomly picked from a set of distances between Europe's 50 largest cities, Warsaw excluded, in such a way that no city appeared twice in any session.

[11] Subjects were allowed to report the interval bounds up to the nearest kilometre but in practice out of 4160 reported values only four were not multiples of 10; vast majority were multiples of 100.

[12] This number was 10'000 in the first session. However, it was subsequently increased because average PST earnings were substantially lower in that session than average CT earnings and quite a few PST subjects were close to bankruptcy. Note that having submitted "agnostic" intervals such as (0,5000), one would only lose 5000 points, half the initial endowment in session 1, over the course of the experiment.

or may not be important empirically). Subject then received joint feedback on the first block the number of questions for which the actual value was below the lower bound, within the interval and above the upper bound. They were also reminded at this point that appropriate 90% confidence intervals should typically yield nine hits in 10 trials.

In the second block of 10 questions, subjects would receive immediate feedback: in the CT they would learn the true value; in the PST, additionally the resulting value of the loss function (when appropriate, also disaggregated into the two components: the one penalizing the width of the interval and the one penalizing having missed the true value) was computed for the subjects.[13]

This design facilitated making a distinction between two alternative hypotheses: that proper scoring leads to (more) truthful reporting directly and that it operates via strengthened feedback. The opposite order of blocks was not considered, because feedback on early intervals would be expected to spill over on subsequent, no-feedback questions. The design admittedly makes it more difficult to distinguish between the impact of feedback and temporal factors such as boredom, but, unless the latter take a rather implausible step-wise form, without any within-block effects, such identification is still possible.

On top of the PST/CT sample split an additional manipulation was also performed. Instructions in the last two sessions were supplemented with a "hint". It read that similar studies in the past and previous sessions of the very same experiment showed that most people give overly narrow intervals, so that on average out of 10 reported intervals only four or five would cover the true value. It was stressed that especially the higher bound is generally too low. Therefore, it was recommended to provide wider intervals than one would intuitively be inclined to do. Similar "cheap talk" interventions were used before (e.g. Block and Harper (1991)).

This additional manipulation was performed between- rather than within-session because we obviously could not truthfully and responsibly give this kind of advice without first obtaining a sufficiently large sample indeed confirming the pre-hypothesized bias for the relevant type of questions, pool of subjects and with or without incentives. Additionally, handling four different treatments in a single session would be somewhat cumbersome (e.g. PST-

[13]The actual distances and resulting losses were not included in the joint feedback after the first block to reduce the likelihood of wealth effects. For example, a subject finding out that she only had 7'000 points left after the first block and aiming to earn at least 20 PLN (requiring 5'000 points) could revert to the strategy of providing intervals of width equal to 4'000 km so that she would be quite sure to lose only 200 points per any of the remaining 10 questions. It was speculated that ignorance concerning current total losses would encourage considering each question separately.

hint subjects' instructions would have been more than twice as long as those of CT-baseline ones).

Introducing the hint served two purposes: firstly, to obtain a sense of how the size of the main treatment effects compares to other interventions. Secondly, to see if perhaps providing incentives to a subject warned about the bias could lead to "overcorrection" (reporting wider intervals than necessary). Clearly, this would speak against the incentivization method.

One hundred and six student subjects participated in the five sessions of the study. Depending on the show-up, the number varied from 16 in session 4 to 23 in sessions 1 and 5. About half majored in economics, 55% were male. Mean age was 22.2.

One of the subjects dropped out having read the instructions (and learned about possible financial losses). One subject dropped out after question no.17 due to other obligations. Both of these were removed from further analysis. Three subjects in the PST treatment went bankrupt but continued playing. One of them made a positive profit in the end, because all participants were given additional 5 PLN due to longer than expected duration of the session. Of the two others, one covered the losses (and vowed never to participate in experiments again). The other refused to do so and the experimenter felt that the amount in question (1.5 euro) did not justify litigation.

Average earnings were about 30 PLN for a session of 40 to 60 minutes.

4 Results

4.1 Treatment effects

Table 1 shows means of key statistics for the two treatments (the entries in the last column will be explained later on).

It can be immediately seen that incentivized subjects submitted wider intervals (with higher upper bounds), and as a results covered the true value more often. It must be noted, however, that the hit rate for the PST was still far below the correct value of .9. Averaging over all choices made by each individual, we obtain individual hit rates. Binomial test rejects the correct calibration hypothesis at 5% level for subjects with less then 16 hits (hit rates of .75 and less). It turns out that *all* CT subjects and 43 out of 51 subjects in the PST fall into this category. No subject reached or exceeded the hit rate of 90%. It should also be noted that the higher hit rate appears to be due to interval width only – the absolute distance between the midpoint of the interval and the actual value ("distance" in table 1) did not differ between the treatments.

Table 1: *Treatment comparison, 2-sided tests of significance*

	Cont. Tr.	Prop. Sc. Tr.	significance
number of subjects	53	51	x
lower bound	976.8	950.9	n.s.
upper bound	1681.2	1838.0	subject, $p < .05$
interval width	704.5	887.1	subject, $p < .05$
hit rate	41.0%	55.0%	subject, $p < .01$
distance	510.9	513.0	n.s.
response time	32.9	40.6	subject, $p < .01$
total loss per period	286.8	256.8	session, $p = .063$

The last row of the table reports the number of points lost per period. Of course, these figures were payoff-relevant for PST subjects only. It turns out that thanks to their wider intervals, the incentivized group lost less points than the control group.

As usual with experiment treatment effects, significance of these results can be tested at three levels. The most conservative approach involves treating each session as just one independent variable. For each variable in Table 1 we have compared the number of cases where the entry for a PST subject was below the median for the session with the number of cases where it was above the median. It is found that PST subjects' total loss in a given period was more often above the session median than below it for each of the five sessions (p value of 2^{-4} or 6.25% for a two sided alternative hypothesis).

The less conservative approach envisages tests which treat choices made by different subjects as independent. Because the experiment involved no interaction between the participants, communication was prevented and session-level effects, if any, were orthogonal to the treatment effects, it seems justified to test at subject level. For this purpose, subject-specific average was calculated for each variable of interest and the treatment effect was tested using the median, ranksum and t tests. The p values for the three tests are quite similar and confirm the results of eye-balling: the impact of treatment on the lower bound of the interval is never significant, the changes in the upper bound and width are significant, at 5% level; the change in the number of penalty points due to a missed true value is significant at 10% level and the change in the number of intervals covering the true value—at 1% level. The results do not change when weights are used to account for slightly unequal assignment to treatments within session (the fraction of incentivized subjects varied from 45.5% to 53.3% due to the odd number of subjects showing up

and due to drop-outs).

The third possibility, taking individual choices as independent observations, does not seem appropriate to test for treatment effects due to possible interdependence of choices made by any individual. We note that even under this approach no impact of the treatment on the lower bound of the interval can be detected.

It was also found that PST subjects took on average approx. 7.7 seconds more to answer each question than CT subjects (on top of taking much more time to read their longer instructions), possibly indicating that they put more effort in providing appropriate values.[14] However, there was no difference in self-reported focus on the task during the experiment.

Regarding the effect of the hint, it could be assessed by comparing behavior in the first three sessions against the last two sessions, as mentioned before. To set the stage for a meaningful test, it was first established that session dummies had no joint significant impact on behavior and there was no trace of a time pattern (which could have resulted e.g. from dissemination of information regarding the nature of the experiment among the participants or selective self-assignment to sessions). It turned out that the hint increased the average width of the interval by some 131 kilometers, thus somewhat less than proper scoring. This effect approaches significance in a one-sided t-test only ($p = .064$), although the power of the test is not comparable to that for the main treatment effect because subjects faced different questions. Also the hit rate increased slightly (from 46.6% to 50.1%, n.s.).

4.2 The role of feedback and learning

As mentioned before, subjects would receive no feedback to the first 10 submitted intervals. Mean hit rates in this block were 32.6% and 50.4% for the CT and PST treatments respectively. When subsequently asked to guess how many of their intervals actually covered the true value, they reported 6.09 and 7.18 on average. Thus, participants correctly recognized that their intervals were too narrow but underestimated the extent to which that was the case. The same pattern was observed before, e.g. in (Sniezek and Buckley 1991). Let us now see how subjects reacted to the information concerning their actual hit rate. Figure 2 shows for each treatment separately, the evolution of the hit rate over time, for each of the five sessions separately and jointly for all sessions.

Several observations can readily be made. First, the hit rate varies substantially between questions. Second, it hardly ever reaches the proper cal-

[14] They may also have spent additional time calculating potential losses.

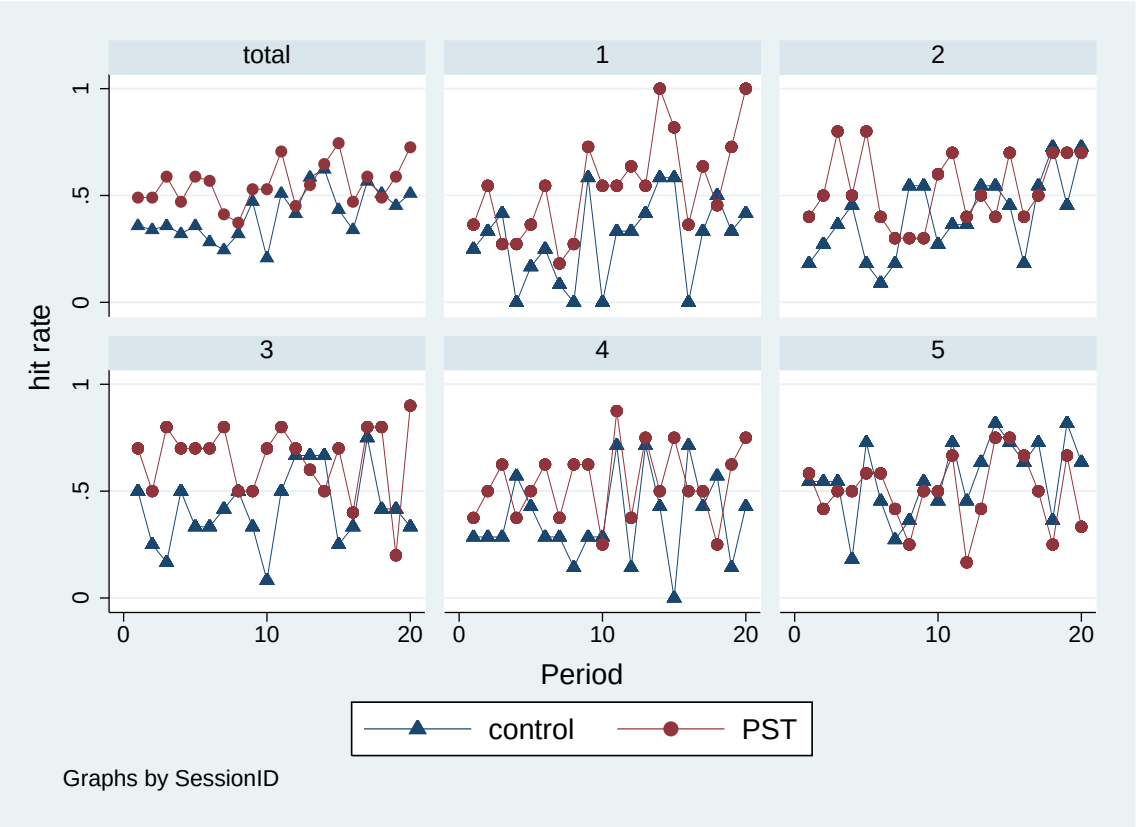

Figure 2: Evolution of hit rate, by treatment and session

ibration level of 90%. Third, there seems to be an upward trend over time, although it may be solely the impact of feedback operating from period 11 onwards. Fourth, except for the level shift (which corresponds to the treatment effect discussed before), the patterns for PST and CT seem quite similar in each session. Finally, also between-session comparison shows rather similar behaviour, except that for some reason the main treatment effect seems weaker in session 5.

To quantify these observations, we have estimated panel logit models explaining whether any given interval covered the true value or not, see table 2.

Model 1 involves fixed effects for subjects and hence time-variant variables only, Model 2 – additionally the main treatment variable and subject characteristics; Model 3 adds interaction between the treatment and block (feedback/no feedback) and period. The significance of the main treatment effect and non-significance of the positive impact of the hint is confirmed. The time course has no direct impact per se, but feedback makes subjects cover the true value more often: the impact is about as strong as that of incentivization. Interactions turn out to be non-significant.

Expectedly, more time spent answering the question leads to higher hit rates. Regarding subject characteristics, individuals who self-declared to have been more focused during the task and males obtained higher hit rates. Notably, males outperformed females in that they covered more true values despite having reported significantly *narrower* intervals. It may have to do with their presumed superior spatial orientation as well as more experience with long-distance driving and planning thereof. Age or academic major were not significant, nor was self-declared willingness to take risk or geographic knowledge.

5 Conclusion

It is widely believed that interval overconfidence is a rather common and persistent feature of human perception. Accordingly, the notion has been extensively applied in behavioral finance and other fields. To date, however, experimental evidence has been solely based on hypothetical questions, raising the natural question of, whether proper incentives can mitigate the bias.

The present study revealed that a simple proper scoring mechanism for confidence intervals is feasible. It may not only be applied in laboratory experiment, but also in the field, wherever verifiable expert interval estimates are sought. For example, options (the strangle strategy) may be used to

Table 2: *Logit regression*

True value covered	Model 1	Model 2	Model 3
PST		.787***	.858***
		[164]	[.265]
Hint		.032	.036
		[.167]	[.167]
Feedback	.884***	.897***	1.152***
	[.195]	[.195]	[.273]
Feedback*PST			-.521
			[.385]
Period	-.014	-.013	-.023
	[.017]	[.017]	[.024]
Period*PST			.018
			[.033]
True value/1000	-.468***	-.493***	-.493***
	[.857]	[.084]	[.084]
Dec. Time (m)	.234	.295*	.282*
	[.152]	[141]	[.142]
Focus		.291***	.290***
		[.076]	[.076]
Risk		.005	.005
		[.064]	[.064]
Geography		.054	.055
		[.069]	[.069]
Econ		.237	.239
		[.175]	[.175]
Male		.414**	.413**
		[.176]	[.176]
Age		.076	.077
		[.066]	[.066]

Model 1 involves fixed effects for subjects, models 2 and 3 – random effects. Standard errors in brackets. Stars indicate significance at 5%, 1% and 0.1% levels.

cheaply provide strong incentives to provide accurate CIs for stock exchange indices. The proposed score may also be calculated ex post, even for non-incentivized CIs, to provide a generic measure of judgment or prediction quality. Additionally, it serves as a simple check regarding information needs of a consumer of an expert opinion – if she feels uncomfortable with the

proposed loss function, then perhaps CIs is not what she needs after all.

The use of the method indeed reduces overconfidence significantly, at least as much as the alternative interventions under scrutiny. Nevertheless, it is by no means able to remove it entirely. While robustness of overconfidence in interval estimation tasks appears to be confirmed, the availability of a proper scoring method now calls for replicating the results that had been obtained using hypothetical questions, comparing i.a. individuals vs. groups, males vs. females and professional vs. novice judges, hard vs. easy, representative vs. unrepresentative questions etc. Similarly, it is worth checking whether disturbing phenomena related to interval judgments, such as insensitivity of interval width to required confidence level (Teigen and Jørgensen 2005) persist when incentives are present. The empirical performance of the straightforward extension of the method to more than two quantiles of the distribution should be assessed.

The proposed method may also be subject to further development. Risk non-neutrality can be dealt with as mentioned before, at a cost of additional time and complexity. Applications to cases where the value to be estimated depends in part on expert's own decisions, such as project leader providing confidence intervals for the project duration, will also require extra care. In particular, punishing overly quick completion as envisaged by the method in its present form will typically be undesirable.

References

Ben-David, I., J. Graham, and C. Harvey (2010): "Managerial miscalibration," Discussion paper, National Bureau of Economic Research.

Biais, B., D. Hilton, K. Mazurier, and S. Pouget (2005): "Judgemental overconfidence, self-monitoring, and trading performance in an experimental financial market," *The Review of economic studies*, 72(2), 287.

Block, R., and D. Harper (1991): "Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis," *Organizational Behavior and Human Decision Processes*, 49(2), 188–207.

Bolger, F., and D. Onkal-Atay (2004): "The effects of feedback on judgmental interval predictions," *international Journal of forecasting*, 20(1), 29–39.

Brier, G. (1950): "Verification of forecasts expressed in terms of probability," *Monthly weather review*, 78(1), 1–3.

Budescu, D., and N. Du (2007): "Coherence and Consistency of Investors' Probability Judgments," *Management Science*, 53(11), 1731.

Connolly, T., and D. Dean (1997): "Decomposed versus holistic estimates of effort required for software writing tasks," *Management Science*, pp. 1029–1045.

Dargnies, M., and G. Hollard (2008): "Monetary Incentives to Learn Calibration: a Gender-Dependent Impact.," .

Fischbacher, U. (2007): "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 10(2), 171–178.

Glaser, M., T. Langer, and M. Weber (2007): "On the trend recognition and forecasting ability of professional traders," *Decision Analysis*, 4(4), 176.

Glaser, M., and M. Weber (2007): "Overconfidence and trading volume," *The Geneva Risk and Insurance Review*, 32(1), 1–36.

Greiner, B. (2004): "The online recruitment system ORSEE 2.0-A guide for the organization of experiments in economics," *University of Cologne, Working paper series in economics*, 10, 2004.

Grinblatt, M., and M. Keloharju (2009): "Sensation seeking, overconfidence, and trading activity," *The Journal of Finance*, 64(2), 549–578.

Grubb, M. (2009): "Selling to overconfident consumers," *The American Economic Review*, 99(5), 1770–1807.

Haran, U., D. Moore, and C. Morewedge (2010): "A simple remedy for overprecision in judgment," *Judgment and Decision Making*, 5(7), 467–476.

Jørgensen, M., K. Teigen, and K. Moløkken (2004): "Better sure than safe? Over-confidence in judgement based software development effort prediction intervals," *Journal of Systems and Software*, 70(1-2), 79–93.

Jose, V., and R. Winkler (2009): "Evaluating Quantile Assessments," *Operations research*, 57(5), 1287–1297.

Kahneman, D., and A. Tversky (1979): "Prospect theory: An analysis of decision under risk," *Econometrica: Journal of the Econometric Society*, pp. 263–291.

Klayman, J., J. Soll, C. González-Vallejo, and S. Barlas (1999): "Overconfidence: It depends on how, what, and whom you ask," *Organizational behavior and human decision processes*, 79, 216–247.

McKenzie, C., M. Liersch, and I. Yaniv (2008): "Overconfidence in interval estimates: What does expertise buy you?," *Organizational Behavior and Human Decision Processes*, 107(2), 179–191.

Murphy, J., and T. Stevens (2004): "Contingent valuation, hypothetical bias, and experimental economics," *Agricultural and Resource Economics Review*, 33(2), 271–281.

Oberlechner, T., and C. Osler (2008): "Overconfidence in currency markets," *Journal of Financial and Quantitative Analysis*.

Offerman, T., J. Sonnemans, G. Van de Kuilen, and P. Wakker (2009): "A Truth Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes*," *Review of Economic Studies*, 76(4), 1461–1489.

Schlag, K., and J. Weele (2009): "Efficient Interval Scoring Rules," *Working Papers (Universitat Pompeu Fabra. Departamento de Economía y Empresa)*, (1176), 1–0.

Smith, V., and J. Walker (1993): "Monetary rewards and decision cost in experimental economics," *Economic Inquiry*, 31(2), 245–261.

Sniezek, J., and T. Buckley (1991): "Confidence depends on level of aggregation," *Journal of Behavioral Decision Making*, 4(4), 263–272.

Soll, J., and J. Klayman (2004): "Overconfidence in interval estimates," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2), 299–314.

Speirs-Bridge, A., F. Fidler, M. McBride, L. Flander, G. Cumming, and M. Burgman (2010): "Reducing overconfidence in the interval judgments of experts," *Risk Analysis*, 30(3), 512–523.

Teigen, K., and M. Jørgensen (2005): "When 90% confidence intervals are 50% certain: On the credibility of credible intervals," *Applied Cognitive Psychology*, 19(4), 455–475.

Van Lenthe, J. (1993): "A blueprint of ELI: a new method for eliciting subjective probability distributions," *Behavior Research Methods*, 25(4), 425–433.

Yaniv, I., and D. Foster (1995): "Graininess of judgment under uncertainty: An accuracy-informativeness trade-off.," *Journal of Experimental Psychology: General*, 124(4), 424.

Appendix: instructions [translated from Polish]

[...] During the experiment you will be asked about distances (**"as the crow flies"**) between pairs of European cities. You will probably not know the exact values, only approximations.

Your task will be to provide in each case an interval covering (containing) the unknown value with a probability of 90%.

In other words, you should be 90% sure that the actual value is between the two numbers you are submitting.

The more precisely you can determine any given distance, the narrower the interval that you should give. Suppose for example that you are being asked about the distance between Zurich and Oslo. If you give a very narrow interval, say between 1050 and 1055 km, you will most probably not cover the actual value (unless it seems to you that you remember this particular number with great precision). The probability that your narrow interval covers the true value is, in your own view, likely much lower than 90%. Submitting a very wide interval, e.g. between 200 and 8000 km, is a mistake, too—(almost) everyone will admit that such an interval will surely (with 100% rather than 90% probability) cover the true distance between Zurich and Oslo. Providing an interval of (700,1350) seems a better idea. Somebody else, more sure about his or her knowledge (and perceiving the distance as larger), can give a narrower interval, e.g. (1200 do 1550).

We would like you to give such an interval that would make the probability that the true value is higher than the upper bound of your interval, as well as the probability that the true value is lower than the lower bound of your interval, each equal to 5%, according to your best judgment.

In other words, you should be 95% sure that the true value is higher than the lower of the two numbers you are reporting and 95% sure that the true value is lower than the higher one.

You are asked to think carefully before each answer.

[what follows was used in the PST only]

To make your thinking pay we will implement a procedure linking your earnings to the intervals you are submitting and the actual

values.

You will start the experiment with 12'000 points. At the end of the experiment each 1000 points will be worth 4 PLN.

You will lose certain number of points with each question. Let us denote the lower and upper bounds of an interval that you submit in response to a question as LB and UB respectively. Let us also denote the actual value as AV. Now, **if it turns out that the interval *does cover* the actual value, that is, $LB \leq AV \leq UB$, you will lose the number of points equal to 5% of the width of your interval in kilometers, i.e. $0.05(UB - LB)$ points.**

If it turns out that your interval *does not cover* the actual value, you will additionally lose the "missing" distance. That is, if $AV > UB$, you will additionally lose $AV - UB$. If $AV < LB$, you will additionally lose $LB - AV$.

Let us consider an example. Suppose that you have submitted the interval (700,1350) for the distance between Zurich and Oslo. Suppose it turned out that the actual value is 900km. That is, we have $LB = 700, UB = 1350, AV = 900$. Your interval covers the actual value. You will lose $0.05(1350 - 700) = 32,5$ points. If it turned out, however, that your interval does not cover the true value, for example, the distance is $AV = 1750$, you would lose much more, namely you will additionally lose the number of kilometers that is missing to cover the true value, in this case, $AV - UB = 1750 - 1350 = 400$, so you will jointly lose 432.5 points. Similarly, if it turned out that the actual value is just 650 km, you would additionally lose $LB - AV = 700 - 650 = 50$, so 83.5 in total.

Note that in order to keep $0.05(UB - LB)$ low, you should give a fairly narrow interval. On the other hand, submitting a narrow interval, you will more often lose points because of not having covered the actual value (and these "penalties", should they occur, will tend to be (much) higher than if you had given a wide interval). The penalty associated with the interval width $UB - LB$ is multiplied by the factor of 0.05. It can be proven that because of this the best thing you can do when trying to maximize your expected payoff is submitting such an interval that you expect the actual value to be lower than your reported lower bound (LB) with a probability of 5% and higher than your reported upper bound (UB) also with a probability of 5%. As a result, your interval will cover the actual value with the remaining probability of 90%. If you want to see the formal proof of this fact, please ask the experimenter.

[information about the timeline of the experiment and about payments—skipped here]

To the best of our judgment, obtaining a negative outcome at the end of

the experiment is very unlikely, as long as you do not take the task too lightly. Should it happen, however, you will have to cover the missing amount with your own money.

If you have any questions, please raise your hand. If everything is clear, please press the button to proceed to the trial periods.

Faculty of Economic Sciences
University of Warsaw
44/50 Długa St.
00-241 Warsaw
www.wne.uw.edu.pl