1.

What Influences Overprecision in Judgmental Forecasting?[1]

Marcin CZUPRYNA[2]
Elżbieta KUBINSKA[3]

Abstract

Previous studies (Krawczyk, 2011; Mannes and Moore, 2013) showed that asymmetric reward functions can be used to get the information on estimated relevant percentiles of the distribution (upper and lower bounds of the confidence intervals, respectively) and thus to analyze the overconfidence level. The estimations obtained by using this indirect method were different than when the participants were directly asked about the value of upper and lower bounds of the relevant confidence intervals. In this article, we consider the problem if these observed differences are permanent and independent of the learning process. In the experiment students provided direct point forecasts and classical lower and upper bounds of the confidence interval and the probability distribution of forecasted weekly rate of returns for WIG and DAX indexes. Based on the reward (loss) functions, indirect estimates of the median (symmetric reward functions) and lower and upper bounds of the confidence interval (asymmetric reward functions) were also derived. There were no significant differences between directly and indirectly provided confidence intervals, implying that the level of overprecision measured by these two methods do not differ if participants are given enough trials to learn the reward function. This suggests studying other than a reward function shape sources of illusion of control. The results have also practical implications, for example for options markets, where the volatility can be estimated directly or indirectly by setting the price of an option (implied volatility).

Keywords: judgmental forecasting, illusion of control, overprecision, asymmetric reward function, confidence intervals

JEL Classification: G4

---

[1] *The authors were supported by a grant awarded by the National Science Centre of Poland under the project title "Behavioral and microstructural aspects of the financial and alternative investments markets", Decision no. 2015/17/B/HS4/02708.*

The authors thank the discussants of the paper and conference participants at the 10th Conference of Polish Academic Association of Economic Psychology for their useful comments and discussions. We also thank the students of the Financial Markets major at the Krakow University of Economics for their participation in the research. In addition, we thank Łukasz Markiewicz for his valuable comments.

[2] *Department of Financial Markets, Faculty of Finance and Law at Crakow University of Economics, Krakow, Poland. Corresponding author. Email: czuprynm@uek.krakow.pl*

[3] *Department of Finance at Krakow University of Economics, Krakow, Poland.*

1. Introduction

Many factors influence decisions under risk and uncertainty. Judgmental forecasting based on subjective perception, contrary to statistical forecasting based on time series models, is an example of such kind of decision-making. Lawrence *et al.* (2006) surveyed the literature and listed the factors that may influence a forecast and its precision. They distinguish two groups of such factors: (1) individual, subjective factors such as: the level of expertise, psychological traits, numeracy skills, or information processing style; and (2) external, objective factors such as the statistical parameters of a predicted time series and the type of forecast. Several forecast types can be distinguished: (a) the point forecast when only a single value is forecast, (b) the probabilistic forecast, when an entire probability distribution function is forecast, and (c) the interval forecast, when lower and upper bounds of the confidence interval are forecast in such a way that the actual value should fall between these bounds with a certain probability. This paper focuses on interval forecasting, where the lower and upper bounds represent the appropriate quintiles of a probability distribution.

This research is motivated from the economic point of view by derivative financial instruments, where investors (1), based on their forecasted distribution of an underlying asset, must fit a proper payoff function taking into account the risks that they want to hedge and (2) estimate the parameter of the distribution of an underlying asset in a direct or indirect way just like in the case of volatility. Within (1) investors must give a price that combines payoff function and risk represented by anticipated distribution of underlying asset (Black and Scholes, 1973; Shimko, 1993). Investors can combine basic derivatives into strategies, like a straddle, where the payoff function is like an absolute value function shifted down for the premiums paid for a put and call option (with the same exercise price). Straddle profits from large price movements, so it is chosen by investors who anticipate high volatility for the underlying asset. If an investor believes that the underlying asset is more likely to decrease than increase, then he/she can take position in strategy that the slope for lower values is double than for higher ones - this is strip strategy. Strip is made by purchasing two puts and one call, so it doubles the potential profits in the case of the downtrend. Strap strategy, which involves two call options and one put, doubles the profits if the underlying asset increases by double slope for values above exercise price. Within options strategies, by taking different combinations of call and put options, investors may hedge against risks by means of asymmetric payoff function they can fit to the skewness of distribution of an underlying asset. Referring to the second point (2), due to the fact that current option prices contain information about future prices that are anticipated by the market, we can find out the implied parameters of the distribution of an underlying asset (Swart and Van Zyl, 2016). The accuracy of volatility parameter estimation is of great importance from a practical and theoretical point of view (Bollerslev, Gibsona and Zhou, 2011). Investors can refer to a direct estimator, *i.e*. standard deviation of historical daily return series or indirect estimator derived from option prices; the relation between those two is of great importance (Christensen and Prabhala, 1998). In this article, we try to look at the process how people estimate the quantiles of the underlying instrument's distribution that affect the payment function and whether the asymmetrical payment improves the estimation in the case of a direct or indirect way of communicating about the distribution.

2. Literature Overview

This paper focuses on overprovision which is one of the forms of overconfidence - a phenomenon commonly observed and reported in the literature. An overconfident forecaster has inaccurate, overly positive perceptions of his/her predictive ability compared to objective criteria (Larrick, Burson and Soll, 2007; Moore and Healy, 2008) In the context of interval forecasts, overconfidence generates an interval that is too narrow compared to the statistically derived forecast confidence interval (Lichtenstein, Fischhoff and Phillips,1982). The obvious explanation would be that humans use the additional knowledge that is not included in the time series. This is postulated by Fama (1970) in the context of different forms of market efficiency, namely: semi-strong and strong). However, this is often rejected based on the empirical evidence that the proportion of actual values of the forecasted variable lying between the given lower and upper bounds of the confidence interval is too low compared to the specified confidence level. The ability to properly estimate risk and to estimate the appropriate quantile is of great importance in finance. The Value at Risk (VaR) measure, commonly used in banking, is defined as a quantile of the loss function. Similarly, binary options and more complex option structures rely heavily on the proper estimation of quantiles.

Many factors may influence the level of overconfidence: (a) statistical properties of time series (Lawrence and O'Connor, 1993), (b) the required confidence level (Önkel and Bolger, 2004), (c) the fact if a point forecast is additionally required or not (Russo and Schoemaker, 1992) and (d) the level of motivation by introducing the monetary and non-monetary pay-offs (Meub, Proeger and Bizer, 2013). Providing the confidence interval in the form of lower and upper bounds may be too abstract. Therefore, the application of an asymmetric loss function was proposed in the literature, *e.g.*, Goodwin (2005) and Lawrence and O'Connor (2005). Mannes and Moore (2013) propose using the asymmetric loss function to analyze the overconfidence level. In their experiment, they asked the participants to predict the temperature for selected past dates and compared the predictions with real temperatures. The reward was based on the difference between each predicted and real temperature. The symmetric and two asymmetric functions were used to calculate a reward (No reward for underestimation, a reward of 5 for overestimation up to 6°C, and a reward of 2 for overestimation of more than 6°C). The problem with such a function is that the prediction results cannot be directly compared with traditional interval estimates. To overcome this, Mannes and Moore (2013) proposed to use the ratio between the observed adjustments (comparing the asymmetric loss function to the symmetric loss function) and the normative adjustments (such adjustments that would maximize the participants' payoff) for this purpose. The adequate reward (loss) function is known and can be applied in quantile regression. Krawczyk (2011) applies such a function in the context of overconfidence estimation.

We use the same function in our research. We designate the *direct interval* to be elicited by directly asking the person for an interval forecast (in the form of lower and upper bounds) such that the real value would fall between the lower and upper bounds with a given probability. We designate the *indirect interval* to be elicited by the application of the appropriate reward (loss) function.

The learning process can significantly influence the accuracy of the estimates obtained in a direct and indirect way. One of the three basic principles of learning is repetition. Two others are active mental processing and making meaningful connections, (see Howe, 1998, pp 7-17). Kim (2015) also presents a three-stage model of learning in the case of complex and ill-structured problems. Therefore, we stipulate that if we allow people to learn how the indirect estimate works, then differences between direct and indirect estimates will diminish.

Numeracy skills and information processing style are subjective factors that may influence the process of learning how to make forecasts. Numeracy is a cognitive characteristic that reveals the level of understanding of numbers, mathematical expressions, and quantitative problem-solving ability (Garofalo and Lester, 1985). Traditional measures of objective numeracy skills evaluate the actual performance of basic numerical problems, *e.g.*, changing percentages into fractions (Lipkus, Samsa and Rimer, 2001; Schwartz *et al*., 1997). The Advanced Berlin Numeracy Test (Cokely *et al*., 2012) is based on fewer (four) items, but they are more advanced tasks. Subjective numeracy skills are measured by using self-report questionnaires and reflect the level of individuals' perceptions of their competences and their preferences for using numbers (Fagerlin *et al*., 2007). Numeracy skills may influence financial decisions, see *e.g.* Donleavy *et al*. (2018) and the learning process itself, also in the context of financial behavior *e.g.* Skagerlund *et al.* (2018). The way that people process information is related to dual-process theories in psychology. (Epstein *et al.,* 1996; Sloman, 1996; Kahneman, 2011). They postulate that human thinking is comprised of two systems. System 1 is devoted to intuitive thinking while System 2 deals with analytical thinking. System 1 is effortless, associative, automatic, and designed for fast answers, while System 2 demands cognitive effort and logical deliberative reasoning. It seems that direct forecasts can be accomplished based only on analyzing charts within System 1. It appears that System 2 needs to be activated to provide accurate, indirect quantile forecasts, since they involve mathematical equations. The ability to learn the proper mechanism of the loss function should be correlated with objective numeracy skills (ONS), but not with subjective numeracy skills (SNS), and with System 2 thinking rather than System 1 thinking.

We first verify the first auxiliary hypothesis, which is an intermediate result, that the complexity of the indirect forecasting method for percentiles elicitation (by the payoff function) requires a learning period for successful performance. Additionally, we check the second auxiliary hypothesis if the rapidity of learning the indirect forecasting method is positively related to numeracy skills. This leads to the following main hypothesis:

H1: The confidence intervals provided by the indirect (by the asymmetric payoff function) forecasting method are similar to those provided by the direct (directly asking for confidence interval lower and upper bounds) forecasting method when the learning period is taken into consideration.

Our paper makes the following claims. First, we will show, contrary to Mannes and Moore (2013) and Krawczyk (2011), that applying the asymmetric reward (loss) function does not influence the estimated level of overconfidence if the participants are given enough time and trials to learn the mechanism of such a function. Also, due to decomposition into two separate problems, the loss function we apply is simpler to understand. The effects observed in the previously cited articles may be due not only to better motivation provided by the asymmetric reward (loss) function compared to the direct questions about the adequate quantiles, but also due to the additional uncertainty induced by the complexity of the asymmetric payoff functions.

3. Method

Participants

Finance students majoring in Capital Markets at a Polish university participated in this experiment during the one-semester (18 weeks) Technical Analysis course. The experiment was carried out on a group of 67 third-year students. We had information on age and gender for 58 students (18 women and 40 men) with the mean age of 22.56 (SD = 3.24). Participation was voluntary; however, participating students were given bonus credits for the Technical Analysis course. Additionally, students with the best results were awarded bonus credits. This was intended to provide higher motivation than any minor monetary payoffs that might have been offered[4]. The data from some students were excluded based on the consistency checks described below and the required number of forecasts provided during the semester.

Procedure

Each week, starting from 1 October 2014 until 8 February 2015, two groups of students were asked to regularly provide the forecasts of the next week's rate of returns for two selected stock market indices: WIG and DAX. The study used an online Lime Survey during the classes to present information, ask the questions, and record the students' answers. During the survey session, the students had online access to historical prices of WIG and DAX indices. Students were told that at the end of the semester the mean absolute deviation from the real observed rates of return and their forecasts would be taken into consideration for the award of bonus credits. Students were asked for the point forecast r_{f} and the lower r_d^L and upper r_d^U bounds of the 80% confidence interval (Students were asked to give values such that the forecasted value will be in the interval with 80% probability.). The lower index d in the symbols r_d^L and r_d^U emphasizes that these estimates of confidence interval lower and upper bounds are obtained by directly asking for them. Students were also asked about the probability distribution for their forecasts. In addition, we asked the students to give two additional forecasts: r_i^L and r_i^U defined by the nonlinear, asymmetric payoff function that is described below in more detail. Analogously, the subscript i emphasizes that these bounds were derived by the indirect method. For r_i^U: if the observed value is above the forecasted value, then the difference between the observed value and the forecasted value is multiplied by 9. Otherwise, if the observed value is below the forecasted value, the forecast error is defined as the difference between the forecasted value and the observed value. For r_i^L: if the observed value is above the forecasted value, then the forecast error is defined as the difference between the observed value and the forecasted value. Otherwise, if the observed value is below the forecasted value, the forecast error is defined as the difference between the forecasted value and the observed value multiplied by 9. In this way, we indirectly asked for the 10% and 90% percentiles, since these statistics minimize the forecast error.

The forecast values, forecast error definitions, and the optimal forecast statistics are summarized in Table 1.

Table 1

Forecast Values, Errors, and Optimal Statistics

Forecast	Forecast error	Optimal forecast
r_f	$\lvert r_f - r_r \rvert$	Expected value
r_i^L	$\begin{cases} 9 \times (r_i^L - r_r)\ if\ r_r < r_i^L \\ 1 \times (r_r - r_i^L)\ if\ r_r > r_i^L \end{cases}$	10th percentile
r_i^U	$\begin{cases} 1 \times (r_i^U - r_r)\ if\ r_r < r_i^U \\ 9 \times (r_r - r_i^U)\ if\ r_r > r_i^U \end{cases}$	90th percentile

[4] *Students receive a monthly stipend depending on their grade point average grade, and a high average allows third-year students to avoid taking the entrance exams for MA studies.*

In this way, we obtained two different forecasts of the 80% confidence intervals: direct (r_d^L; r_d^U) and indirect (r_i^L; r_i^U).

We first asked for WIG forecasts and then for DAX forecasts. The ordering of questions was first the point forecast r_f, then the direct confidence interval (r_d^L; r_d^U) and probability distributions forecast, and finally the indirect confidence interval (r_i^L; r_i^U).

Scales measuring subjective and objective numeracy skills were used. Objective numeracy skills were measured by the 11-item Objective Numeracy Scale (ONS) proposed by Lipkus, Samsa and Rimer (2001). The result of scoring this scale is the number of correct answers to the questions, so the bigger the values of ONS, the better are the numeracy skills. Students also completed the 8-item Subjective Numeracy Scale (SNS) proposed by Fagerlin *et al*. (2007) to measure perceived numeracy. SNS is a self-report measure of the perceived ability to perform various mathematical tasks and preferences for the use numerical versus prose information. This questionnaire has two subscales: (1) cognitive abilities (SNS A) and (2) preference for display of numeric information (SNS B). The scores assigned to SNS (A) and SNS (B) are the means of the values given by the respondents to the questions within the subscales. The 3-item Cognitive Reflection Test (CRT), as proposed by Frederick (2005) was also used. The bigger the CRT score, the more likely it is that the individual can suppress the initial pressure to make an incorrect System 1 "gut" response and override it with more reflective System 2 thinking to make a correct response.

4. Results

It was possible for each of 67 students to provide 18 rounds of forecasts. Participation was less than perfect; we gathered 915 of weekly forecasts in total, leading to the mean number of forecasts provided was 13,66 (SD = 4,44). Each data point contained point and interval forecasts for WIG and DAX index return rates. This database was supplemented by numeracy skills and CRT data from online surveys.

We tested the first auxiliary hypothesis that the complexity of the indirect forecasting method for percentiles elicitation (by the payoff function) requires a learning period for successful performance.

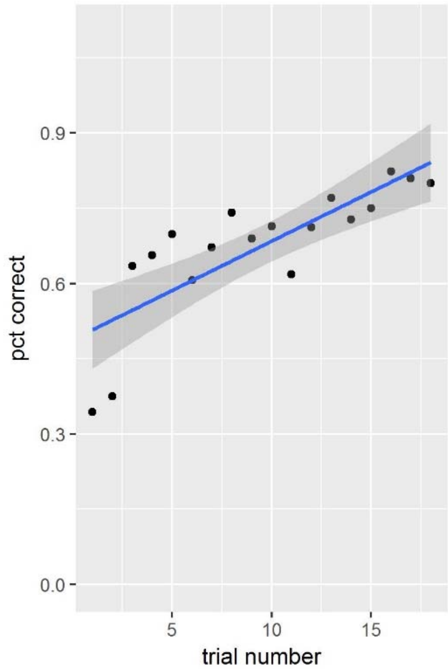
For this purpose we analyzed how quickly the students had understood the mechanism behind the proposed loss function and had answered in a coherent way. For this purpose, we defined two coherency conditions. These conditions stipulate that the ordering of the point forecast and 10% and 90% percentiles is properly in order if:

$$r_i^L < r_f < r_i^U.$$

We evaluated the percentage of errors in each of the forecasts at the group level and at the individual level (Figure 1). We cannot compare these measures of coherency for indirect forecasting with similar measures for the direct confidence intervals. For the latter case, the coherence was supported by the LimeSurvey data.

Figure 1

Percentage of Coherent Forecasts of Weekly WIG and DAX Rates of Return (Both Must Be Coherent at the Forecast Date) for Each Forecast round



We can observe that the percentage of coherent forecasts is low at the beginning of the semester (around 35%) and gradually increases to 80%, which qualitatively confirms.

These qualitative observations are supported by the regression results of the percentage of coherent forecasts in the number of the forecast round. The regression coefficient equals 0.02 (each round number percentage of coherent forecasts increases by 2% on average) and is statistically significant, see Table 2. This suggests that asymmetric pay-off functions (as in the case of e.g. options) may cause problems in understanding, and require a relatively longer learning period.

Table 2

Regression Results

	Percentage of coherent forecasts		
Predictors	Estimates	CI	P
(Intercept)	0.49	0.41 – 0.57	<0.001
number trial	0.02	0.01 – 0.03	<0.001
Observations	18		
R^2 / adjusted R^2	0.641 / 0.619		

We also analyzed the forecast series for each student individually and identified the “learning round”, *i.e.*, the first forecast that is followed only by coherent forecasts (the moment when the student understood the mechanism and provided coherent forecasts from that moment onward). We considered each index separately. In addition, to accept a student’s data set, we required the student to make at least 5 more coherent forecast rounds after the learning round. However, only 30 (out of 67) students satisfied this condition. Therefore, we relaxed this condition by allowing for one incoherent forecast round after the learning round. This relaxation increased the number of students providing data to 36. The variables for the number of the learning round for the indices are denoted by LR(WIG) and LR(DAX), respectively. The distributions of the learning rounds for WIG and DAX are shown in Figures 2 and 3.

Figure 2

The Distribution of the Learning Rounds for WIG Index - LR(WIG), with Tolerance Condition of One Error

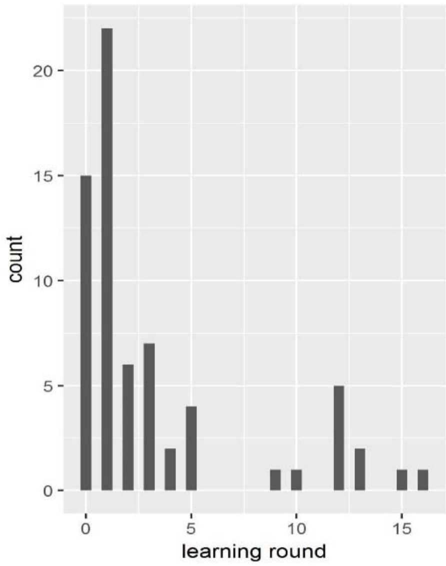
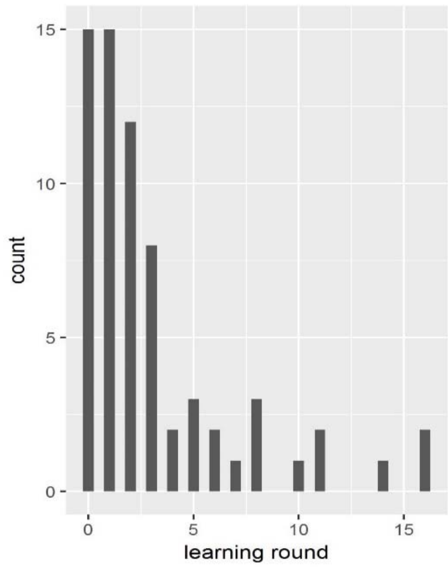


Figure 3

The Distribution of the Learning Rounds for DAX Index - LR(DAX), with Tolerance Condition of One Error



There were 15 of the 67 students that never learned how to provide the coherent indirect forecasts for either index (learning round is set to 0 in this case in Figures 2 and 3). The other students started giving the coherent forecast since 6th round in the case of DAX and WIG index (for DAX mean=6,04, SD = 4,94 and for WIG mean=5,64, SD = 5,07).

We also test the second auxiliary hypothesis that the rapidity of learning the indirect forecasting method is positively related to numeracy skills.

We tested the relationship between the learning round and numeracy skills and CRT scores. Table 3 generally shows negative correlation coefficients between measures of numeracy skills, processing information system, and learning round, which suggests that the higher the numeracy skill, the sooner students learned to properly adjust their forecasts to the nonlinear asymmetric payoff functions.

Table 3

Correlation Coefficients between Learning Round (The Round That A Participant Had Learned How to Forecast) and Subjective Numeracy Test, Objective Numeracy Test and CRT test (N=40)

	LR(DAX)	LR(WIG)	SNS (A)	SNS (B)	ONS	CRT
LR(DAX)	1	0.71	-0.08	-0.16	0.01	-0.22
LR(WIG)	0.71	1	0.05	-0.26	-0.08	-0.22
SNS (A)	-0.08	0.05	1	0.11	0.35	0.25
SNS (B)	-0.16	-0.26	0.11	1	0.19	0.49
ONS	0.01	-0.08	0.35	0.19	1	0.34
CRT	-0.22	-0.22	0.25	0.49	0.34	1

Table 4

P Values For Correlation Coefficients between Learning Round (The Round That A Participant Had Learned How to Forecast) and Subjective Numeracy Test, Objective Numeracy Test and CRT test (N=40)

	LR(DAX)	LR(WIG)	SNS (A)	SNS (B)	ONS	CRT
LR(DAX)		0	0.62	0.31	0.94	0.18
LR(WIG)	0		0.76	0.1	0.62	0.17
SNS (A)	0.62	0.76		0.5	0.03	0.12
SNS (B)	0.31	0.1	0.5		0.23	0
ONS	0.94	0.62	0.03	0.23		0.03
CRT	0.18	0.17	0.12	0	0.03	

However, there are no significant relationships between ONS, SNS and CRT scores and the learning speed measured by the number of the learning round. Second auxiliary hypothesis did not receive statistical support.

The joint effect of numeracy skills on learning speed (measured by the learning round, the higher the learning round the lower the learning speed) is tested using multivariate regression analysis, see Tables 5 and 6. No significant effects were found. Neither subjective nor objective numeracy skills influenced the speed of the learning process.

Table 5

Regression Results of Learning Round (WIG) on Numeracy Skills

	Learning round (WIG)		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	20.22	3.44 – 37.00	0.024
SNS (A)	-0.15	-1.63 – 1.33	0.843
SNS (B)	-1.34	-4.31 – 1.64	0.384
ONS	-8.50	-23.17 – 6.16	0.264
CRT	-3.39	-7.85 – 1.08	0.146
Observations	40		
R^2 / adjusted R^2	0.223 / 0.134		

Table 6

Regression Results of Learning Round (DAX) on Numeracy Skills

	Learning round DAX		
Predictors	*Estimates*	*CI*	*P*
(Intercept)	13.50	-0.61 – 27.61	0.069
SNS (A)	-0.59	-1.83 – 0.65	0.358
SNS (B)	-0.72	-3.22 – 1.78	0.577
ONS	-2.73	-15.07 – 9.60	0.667
CRT	-2.20	-5.95 – 1.55	0.259
Observations	40		
R^2 / adjusted R^2	0.147 / 0.050		

We test the main hypothesis **H1**: The confidence intervals provided by the indirect (by the asymmetric payoff function) forecasting method are similar to those provided by the direct (directly asking for confidence interval lower and upper bounds) forecasting method when the learning period is taken into consideration.

To verify H1, the lengths of the students' confidence intervals were calculated, based on their answers: r_d^U, r_d^L, r_i^U, and r_i^L.

- Variant1: the difference between direct estimates of quantiles: $r_d^U - r_d^L$,
- Variant2: the difference between indirect estimates of quantiles: $r_i^U - r_i^L$,

The intervals were normalized by dividing them by the length of the confidence interval calculated based on estimated parameters of GARCH (1,1) models for returns. Data were used only from the coherent forecasts starting from each student's learning round for the direct and indirect methods. The results are presented in Tables 7 and 8 for WIG and DAX indices, respectively. The second rightmost column presents the p-value results of testing the paired t test. Similar results were also observed when testing the mean rank differences by the Wilcoxon rank sum test, (Bauer, 1972; R Core Team, 2016). The rightmost column presents the results of the power test with effect size (Cohen's d - which is the difference between the means divided by the pooled standard deviation) value of 0.5, (Cohen, 1988; Champely, 2018).

Table 7

Comparison of Confidence Intervals' Lengths for WIG Index for Variant 1 (1) and Variant 2 (2) for Individual Rounds

trial	N	Mean (1)	Sd(1)	Mean(2)	Sd(2)	P value	Power
1	20	1.23	0.63	0.98	0.57	0.3	0.88
2	23	1.03	0.5	0.95	0.48	0.41	0.94
3	29	0.9	0.37	0.9	0.46	0.66	0.99
4	31	0.9	0.48	0.95	0.45	0.75	0.99
5	34	0.83	0.35	0.84	0.38	0.63	0.99
6	33	0.81	0.35	0.83	0.46	0.99	1
7	33	0.87	0.4	0.9	0.4	0.16	0.92
8	33	0.77	0.34	0.83	0.42	0.3	0.97
9	33	0.77	0.32	0.85	0.4	0.9	1
10	32	0.76	0.35	0.88	0.42	0.22	0.94
11	32	0.92	0.43	1.05	0.48	0.6	0.99
12	33	0.9	0.4	0.99	0.42	0.32	0.97
13	33	0.81	0.36	0.82	0.37	0.72	0.99
14	30	1.02	0.4	1.07	0.46	0.69	0.99
15	28	1.12	0.48	1.2	0.58	0.73	0.99
16	26	1.17	0.56	1.18	0.68	0.47	0.97
17	17	1.05	0.5	1.21	0.58	0.2	0.77
18	4	0.66	0.36	0.84	0.48	0.2	0.35

Table 8

Comparison of Confidence Intervals' Lengths for DAX Index for Variant 1 (1) and Variant 2 (2) – for Individual Rounds

trial	N	Mean (1)	Sd(1)	Mean(2)	Sd(2)	P value	Power
1	15	0.95	0.44	0.88	0.58	0.73	0.96
2	22	0.93	0.63	1.62	3.65	0.38	0.93
3	29	0.86	0.43	0.81	0.46	0.32	0.95
4	29	0.82	0.46	0.87	0.5	0.3	0.95
5	32	0.74	0.39	0.76	0.41	0.59	0.99
6	32	0.69	0.4	0.69	0.33	0.96	1
7	32	0.69	0.36	0.75	0.39	0.18	0.93
8	34	0.67	0.31	0.7	0.39	0.51	0.99
9	34	0.67	0.25	0.65	0.32	0.61	0.99
10	34	0.66	0.32	0.72	0.33	0.24	0.96
11	34	0.78	0.42	0.73	0.3	0.4	0.98
12	34	0.79	0.35	0.81	0.42	0.56	0.99
13	33	0.7	0.34	0.68	0.36	0.72	0.99
14	30	0.71	0.34	0.72	0.3	0.69	0.99
15	28	0.74	0.31	0.76	0.33	0.73	0.99
16	26	0.81	0.4	0.76	0.42	0.47	0.97
17	17	0.75	0.35	0.84	0.39	0.2	0.77
18	4	0.5	0.16	0.77	0.31	0.2	0.35

The results presented in Tables 7 and 8 does not allow to reject hypothesis H1, that there are no significant differences in length of direct and indirect confidence intervals, which implies no differences in overprecision bias measured by the indirect and direct methods.

With reference to the analogy with the option theory, no significant differences between implied and directly estimated (using all the available information) should be observed in the long term.

We have additionally validated this result. Namely, we excluded all the observations, such that the conditions are satisfied $r_i^L = r_d^L$ and $r_i^U = r_d^U$ from the analysis (17,03% for DAX forecasts and 21,03% of WIG forecasts). This conservatively excludes the cases when the students might have had learned, based on experience or external knowledge, that 10% and 90% quintiles are optimal solutions to the problem of the minimization of the expected loss in case of the applied loss functions (indirect method). The results comparing non-normalized intervals for direct (1) and indirect (2) methods are presented in Table 9.

Table 9

Restricted Comparison of Direct and Indirect Confidence Interval Lengths for WIG and DAX Indices

	WIG	DAX
Mean (1)	0.04	0.04
Sd (1)	0.01	0.01
Mean (2)	0.04	0.05
Sd (2)	0.01	0.02

The p-values of the Wilcoxon test on equality of group mean ranks (Bauer, 1972) are 0.58 for the WIG index and 0.98 for the DAX index. The differences between direct and indirect intervals are not significantly different, even after exclusion.

Similar results were obtained in a second experiment, where 70 students participated in weekly forecasts from 10 October 2016 until 30 January 2017. The scope of the second experiment was the same as the first experiment. All the experiment results were similar and therefore not presented in the paper.

5. Conclusions

We have shown that there are no statistically significant differences between direct and indirect confidence intervals and thus the measured level of overprecision, provided the participants are given enough time and trials to learn the mechanism of such a function. In such a way, contrary to previous studies (Mannes and Moore, 2013; Krawczyk, 2011), we found that the form of confidence intervals elicitation (direct question versus indirect method with payoff function) does not influence the length of the intervals in a systematic way. As a group, the participants made fewer mistakes in every subsequent round of the experiment. Individual characteristics did not influence the rapidity of learning how to properly provide forecasts with nonlinear asymmetric payoff functions.

There are limitations to this study. First, students were the only type of participants. Second, both methods, direct and indirect, have only been introduced in the experiment and explained in one way. Third, the learning process was extended for weeks. The limitations of this experiment can be an inspiration for further research. Extending the type of participants and using different asymmetric reward functions as well as different ways of introductions of these methods, e.g. different forms of (graphical) presentation, will be the subject of further research. Carrying out an analogous study in the financial derivatives (options) context would also be of interest.

Additionally, we noticed some indications that overprecision, as measured by the length of confidence intervals, may be due to the carry-over effect of the risk estimates from a well-known time series, like the WIG index for Polish students, to one that is less familiar, the German DAX index. This phenomenon would encourage research in the Bayesian approach to overprecision level measurements. Namely, we should verify and take into consideration the prior probabilities for confidence levels of estimates (*e.g.*, derived from a comparable well-known time series). This would also be the subject of the further research.

References

Bauer, D.F., 1972. Constructing confidence sets using rank statistics. *Journal of the American Statistical Association*, 67(339), pp.687-690.

Black, F. and Scholes, M., 1973. The Pricing of Options and Corporate Liabilities, *Journal of Political Economy.* 81 (3), pp. 637–654.

Bollerslev, T., 1986. Generalized Autoregressive Conditional Heteroscedasticity, *Journal of Econometrics,* 31, pp.307–327.

Bollerslev, T., Gibson M. and Zhou, H., 2011. Dynamic Estimation of Volatility Risk Premia and Investor Risk Aversion from Option-Implied and Realized Volatilities, *Journal of Econometrics*, 160(1), pp. 235–245.

Champely, S., 2018. pwr: Basic Functions for Power Analysis. R package version 1.2-2., Available at: <https://CRAN.R-project.org/package=pwr>.

Christensen, B.J. and Prabhala, N.R., 1998. The relation between implied and realized volatility, *Journal of Financial Economics,* 50, pp.125-150.

Cokely, E.T. *et al*., 2012. Measuring risk literacy: The Berlin Numeracy Test. *Judgment and Decision making.* 7(1), pp.*25–47.*

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences*. NY: Routledge.

Donleavy, G.D. *et al*., 2018. How numeracy mediates cash flow format preferences: A worldwide study. *The International Journal of Management Education*, 16(2), pp.180-192.

Epstein, S., Pacini, R., Denes-Raj, V. and Heier, H., 1996. Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of personality and social psychology*, 71(2), 390.

Fagerlin, A. *et al*., 2007. Measuring numeracy without a math test: Development of the Subjective Numeracy Scale (SNS). *Medical Decision Making*, 27, pp.672-680.

Fama, E.F., 1970. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25 (2), pp. 383-417.

Frederick, S., 2005. Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4), pp. 25-42.

Garofalo, J. and Lester, F. K., 1985. Metacognition, cognitive monitoring, and mathematical performance. *Journal for research in mathematics education*. 16(3), pp. 163-176.

Goodwin, P., 2005. Providing support for decisions based on time series information under conditions of asymmetric loss. *European Journal of Operational Research,* 163(2), pp.388– 402.

Howe Michael J.A., 1998. *Principles of abilities and human learning*, London: Psychology Press Ltd (Taylor & Francis group).

Hyman, J. M., 1983. Accurate monotonicity preserving cubic interpolation. *SIAM Journal of Scientific Statistical Computing,* 4, pp.645–654.

Kahneman, D., 2011. *Thinking, fast and slow*. New York: Macmillan.

Kim, M.K., 2015. Models of learning progress in solving complex problems: Expertise development in teaching and learning. *Contemporary Educational Psychology*, 42, pp.1-16.

Krawczyk, M., 2011. Overconfident for real? The proper scoring for confidence intervals, University of Warsaw Faculty of Economic Sciences, Working papers 2011/15.

Lawrence, M. and O'Connor, M., 1993. Scale, randomness and the calibration of judgmental confidence intervals. *Organizational Behavior and Human Decision Processes*, 56, pp.441– 458.

Lawrence, M. and O'Connor, M., 2005. Judgmental forecasting in the presence of loss functions. *International Journal of Forecasting*, 21, pp.3–14.

Lawrence, M., Goodwin, P. O'Connor, M. and Önkal, D., 2006. Judgmental forecasting: A review of progress over the last 25 years*, International Journal of Forecasting*, 22, pp.493– 518.

Lichtenstein, S., Fischhoff, B. and Phillips, P.,1982. Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, and A. Tversky, Eds. *Judgment* *under uncertainty: Heuristics and biases* (pp.306– 334). Cambridge University Press.

Lipkus, I. M., Samsa, G. and Rimer, B. K., 2001. General performance on a numeracy scale among highly educated samples. *Medical decision making*, 21(1), pp. 37-44.

Mannes, A.E. and Moore, D.A., 2013. A behavioral demonstration of overconfidence in judgment. *Psychological Science*, 24(7), pp.1190-1197.

Meub, L., Proeger, T. and Bizer, K., 2013. Anchoring: a valid explanation for biased forecasts when rational predictions are easily accessible and well incentivized? Center for European Governance and Economic Development Research, *Discussion Paper* 166-2013.

Önkal, D. and Bolger, F., 2004. Provider-user differences in perceived usefulness of forecasting formats. *Omega*, 32, pp.31– 39.

R Core Team, 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna. Austria. Available at: <https://www.R-project.org/.>

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

Schwartz, L.M., Woloshin, S., Black, W.C. and Welch, H.G., 1997. The role of numeracy in understanding the benefit of screening mammography. *Annals of internal medicine*, 127(11), pp. 966-972.

Shimko, D.C., 1993. Bounds of probability, *Risk (Concord, NH)*, 6 (4), pp. 33-37.

Skagerlund, K., Lind, T., Strömbäck, C., Tinghög, G. and Västfjäll, D., 2018. Financial literacy and the role of numeracy–How individuals' attitude and affinity with numbers influence financial literacy. *Journal of behavioral and experimental economics*, 74, pp. 18-25.

Sloman, S.A., 1996. The empirical case for two systems of reasoning. *Psychological bulletin*, 119(1), pp. 3-32.

Swart, B. and Van Zyl, J., 2016. Recovery of asset information from options prices, *Investment Analysts Journal, 45* (2), pp. 110-122.