Eliciting interval beliefs: An experimental study*

Ronald Peeters† Leonard Wolk‡

January 21, 2015

*** Work in progress ***

Abstract

In this paper we study the use of the interval scoring rule as a non-market based forecasting mechanism. In our experiment subjects forecast the termination time of a time series to be generated from a given but unknown stochastic process, where over time they gradually learn more about the underlying process and hence the true distribution over termination times. We conduct two treatments, one with a high and one with a low volatility process. We find that individuals forecast better when facing a low volatility process, but when individual forecasts are aggregated over groups, groups make better predictions when facing a high volatility process.

JEL Classification: C53, D83, C91.

Keywords: forecasting, belief elicitation, learning, experiment.

*We thank the audiences at the Borsa Istanbul 2014 Workshop on Behavioral Finance, the 2014 ESA meetings in Prague and Fort Lauderdale, the INFORMS Annual Meeting 2013 in Minneapolis, the Maine Economics Conference 2014 in Waterville, the University of Paderborn, as well as Matt Embrey, Ben Gillen, Jörg Gross, Glenn Harrison, Stephan Smeekes, Sasha Vostroknutov, Joël van der Weele and Maria Zumbühl. Ronald Peeters gratefully acknowledges funding from NWO. A previous version of this paper circulated with the title "Eliciting and aggregating individual expectations: An experimental study".

†Department of Economics, Maastricht University. E-mail: r.peeters@maastrichtuniversity.nl

‡Department of Economics, Colby College. E-mail: leonard.wolk@colby.edu

1 Introduction

Firms often depend on internally generated forecasts when making operational decisions such as whether to invest in a project or whether to increase production capacity. Generating forecasts for such purposes require both the elicitation of beliefs and aggregation of information, dispersed across different individuals within, as well as outside, a firm. Given that unstructured mechanisms to aggregate information may result in a failure to correctly take all information into account (Hopman, 2007), it is important to investigate the forecasting ability of alternative mechanisms designed to elicit beliefs.

In this paper we propose and implement a non-market based mechanism to elicit forecasts and to test it experimentally. Non-market based methods have recently been shown to perform well and Gillen et al. (2013) implement such a mechanism to forecast future sales within Intel and are able to outperform internal forecasts in a majority of the cases. Goel et al. (2010) find that non-market based methods do not perform significantly worse than prediction markets in forecasting outcomes of sports and movie events. Prediction markets have been extensively studied as an elicitation mechanism and they have been shown to perform very well in a wide array of applications, such as elections (Forsythe et al., 1992), corporate settings (Cowgill and Zitzewitz, 2013) and in the laboratory (Smith, 1962). Yet, these markets do not come without problems and can be subject to manipulations (Hanson et al., 2006; Veiga and Vorsatz, 2006) as well as strict regulatory requirements (Arrow et al., 2008).

In our experiment subjects have to forecast, over a sequence of twenty periods, the termination time of a time series that is to be generated from a fixed but unknown random process by specifying an interval where they believe the time series is going to terminate. Subjects are not informed about the details of the process and gradually learn about the underlying parameters as the experiment advances. One advantage of conducting a laboratory experiment is that we are able to control the distribution over possible outcomes (the random process being just one way to generate such a distribution). This allows us to compare the individual predictions against the true ex-ante distribution of outcomes – something that is hard to do with experiments in the field where comparison are typically made against realizations – and hence facilitates a better performance analysis and allows a better understanding in the mechanics that lead to good forecasting environments.

We incentivize subjects by means of the interval scoring rule (Schlag and van der Weele, 2009). That is, only a positive payoff is earned in case the realized termination time is in the stated interval – this payoff being decreasing in the chosen length of the interval – and zero otherwise. Consequently, the experimental setting does not involve strategic interaction, i.e. there is no competition among subjects and they are rewarded purely on basis of their own performance. There are several advantages of moving away from a market-based setting. For instance, non-market based mechanisms can be operated with fewer forecasters, aggregation of

individual forecasts can be weighted by individual characteristics of the forecaster (including past performance) and information flows can easily be traced across different subsets of the forecasters.

Several papers have implemented variations on the interval scoring rule, including Galbiati et al. (2013), Tausch et al. (2014) and Peeters et al. (2012); yet its properties have not been studied extensively. In this paper we explore, in a forecasting context, several aspects of belief elicitation using this rule. First, we consider the choices individuals make in this environment given the incentives provided, and how choices change over time in response to recent experiences. Second, we study how individual forecasting performance relates to the level of the underlying uncertainty and individual attributes like cognitive ability, risk and gender. Finally, we investigate the quality of the performance of group predictions (on basis of aggregating the forecasts of the individuals in this group) depending on the size and composition of the group.

We find that individuals' forecasts are significantly better in the low volatility treatment than in the high volatility treatment. Over time individuals improve their forecasting performance in the low volatility treatment, but fail to do so significantly in the high volatility treatment where they learn to improve in the choice of location of the interval given the interval length, but fail to choose the correct length. Interestingly, behavior, as well as performance, in the experiment does not appear to be significantly affected by risk preferences. This is in line with Harrison et al. (2013), who show that when eliciting subjective beliefs over continuous events using a popular scoring rule one does not need to correct those beliefs for the subject's risk preferences. Yet, interestingly, this is in contrast to beliefs elicited over binary outcomes which are affected by an individual's risk tolerance (see for instance Winkler and Murphy, 1970).

When aggregating forecasts over groups of individuals, we find that the group performance (as measured by the Hellinger distance to the true distribution) is increasing in group size at a decreasing rate. While, for any given group size, group performance is better in the low volatility treatment throughout the first half of the experiment, aggregated forecasts are better in the high volatility treatment during the second half. This is possibly due to there being less correlation in individual forecasting errors, which makes the aggregate forecast resemble the underlying distribution better. Although we believe we may conclude that the mechanism studied yields a quite good forecast already when aggregating over few individuals, forecasting accuracy can be improved when aggregating over the right individuals. For instance, Budescu and Chen (2014) and Goldstein et al. (2014) show that performance can be improved by putting more weight on individuals that performed better in the past. Our results show that groups perform better when the share of females is larger and the average tolerance towards risk is higher; the effect of cognitive ability seems to interact with the level of uncertainty.

2 Experiment

In the experiment subjects are exposed to a random process that starts at a value of zero at time $t = 0$ and runs from there in discrete time-steps. At each unit of time the value is incremented with a real number (possibly negative) that is drawn randomly according to a normal distribution with mean zero (hence, there is no drift) and a fixed but unknown variance. The process terminates either when the value crosses the lower boundary at -2.5, crosses the upper boundary at $+2.5$, or has reached time $t = 100$ without having reached one of these boundaries. Figure 1 shows one time series generated by this process that led to a termination at the lower bound at time $t = 63$. In a sequence of twenty rounds, the task of the participants in this experiment is to predict the termination time of the upcoming time series. While doing so, the participants gradually learn about the underlying parameters that generate the stochastic process, possibly giving rise to a gradual improvement in their predictions.

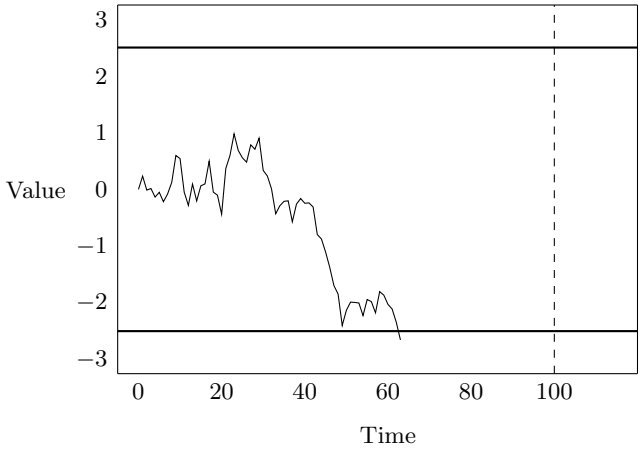


Figure 1: An example of a time series.

Prior to the first round, participants were shown an animation of a randomly generated time series. After having seen this animation, they were asked to indicate the time interval in which they believe the next time series is going to hit one of the boundaries, conditional on the time series to terminate before time $t = 100$. The decision was made by positioning two triangular cursors along the time line between $t = 0$ and $t = 100$.[1] Participants were incentivized by means of an interval scoring rule (Schlag and van der Weele, 2009): a participant expressing the belief that, conditional on the time series to terminate before time $t = 100$,

[1] In addition to this task described, during all twenty rounds, the subjects were simultaneously confronted with a second decision task. In this second task subjects were asked how likely they regard the event that the time series will terminate before $t = 100$. As this refers to a binary event, for this task they were incentivized by means of a quadratic scoring rule. This task was implemented for the sole purpose to provide subjects with the full set of outcomes space, but their decisions for this task are not subject to analysis in this paper.

it to hit one the boundaries within the time interval $[\hat{x}, \hat{y}]$ received $100 \cdot (1 - \frac{\hat{y}-\hat{x}}{100})^2$ ECU (Experimental Currency Units) if the time series indeed terminated within the given time interval and received nothing otherwise. Thus, the payoff that could potentially be obtained is larger when a smaller interval is selected and the potential payoff was shown on-screen in real-time while cursors were moved along the time line. After having confirmed their predictions, participants were shown the animation of the time series that was generated for the first round, whereafter the task was repeated in the second round. This procedure continued until the last (twentieth) round.

Finally, the participants participated in a short cognition task in which we elicited their perceptual reasoning ability, their risk attitude, and a few personal characteristics, including gender and age. For the cognition task, we used the symbol-digit correspondence test from the Wechsler Adult Intelligence Scale (WAIS), in which subjects had 90 seconds to find as many correspondences between symbols and numbers as they could, using the correct number for each symbol. The speed and accuracy of this task under time pressure determine an individual's perceptual reasoning ability (cf. Dohmen et al., 2010). Risk attitude was elicited by the direct approach as suggested in Dohmen et al. (2011).

A random selection of subjects from our subject pool (mainly students in business and economics) were invited to participate in one of two sessions of an economic experiment via ORSEE (Greiner, 2004). Both sessions were run in the BEElab at Maastricht University in September 2013. The instructions were paper-based and the prediction phase was computerized using z-Tree (Fischbacher, 2007).[2] In total 48 students participated: half of them participated in the low volatility treatment with the standard deviation of the normal distribution being equal to 0.1885, the other half participated in the high volatility with this standard deviation being set at 0.2270.[3] All participants in a treatment were shown the same animations in the same order, and the series of time series were generated by a statistical software package and were not subject to experimental manipulation. At the end of the session, for each participant individually, eight random draws (with replacement) over the payoffs that were earned in the twenty rounds were made. The final earnings of the participants consisted of the amount of ECUs collected in these eight tasks exchanged into Euros at a conversion rate of 6 Eurocents for each ECU and a 3 Euro show-up fee. Each experimental session lasted about 60 minutes and the average earnings of the subjects was 13.56 Euro.

Figure 2 presents the true distribution over termination times, conditional on termination before $t = 100$, for the two treatments. The mode of this distribution is at 66 for the low volatility treatment and at 31 for the high volatility treatment. Given the incentives provided, when having perfect knowledge of this true distribution, a risk neutral individual maximizes

[2]The instructions as they were provided to the experimental subjects are included in Appendix A.

[3]These standard deviations are chosen such that the probability of the process to terminate before $t = 100$ equals approximately 1/3 in the low volatility treatment and 2/3 in the high volatility treatment.

her expected payoff by choosing the interval $[51, 83]$ in the low volatility treatment and the interval $[21, 51]$ in the high volatility treatment.

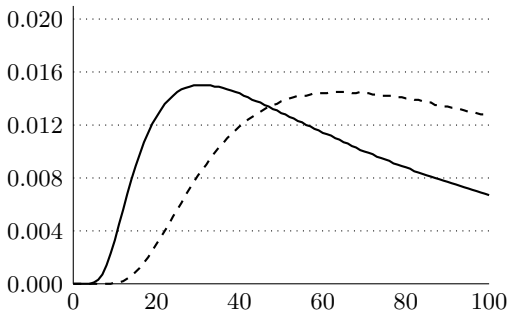


Figure 2: Distribution over termination times conditional on termination before $t = 100$. The dashed curves relate to the low volatility treatment; the solid curve to the high volatility treatment.

3 Results

In Table 1 we present the summary statistics of our experiment. The upper part shows the summary statistics of the main characteristics of the participants in our experimental sessions. The ratio of males was slightly larger in the low volatility treatment; so was the number of correctly identified symbols in the cognition task. There are no substantial difference in age and risk attitude (where the value 0 indicates extreme risk aversion and the value 10 extreme risk loving) between the participants in the two treatments.

	Mean value (std.dev)					
	All		Low		High	
Age (years)	21.2	(2.4)	21.1	(1.9)	21.2	(2.8)
Gender (%, Male = 1)	50.0%		58.3%		41.7%	
Risk attitude (0–10)	6.1	(1.9)	6.0	(1.9)	6.1	(2.0)
Cognitive ability (number)	40.5	(6.5)	41.1	(7.3)	40.0	(5.6)
Lower bound (0–100)			43.6	(13.5)	31.6	(15.6)
Upper bound (0–100)			82.4	(12.9)	77.1	(15.8)
Exp. payment (in ECU)			17.6	(3.8)	14.1	(3.6)

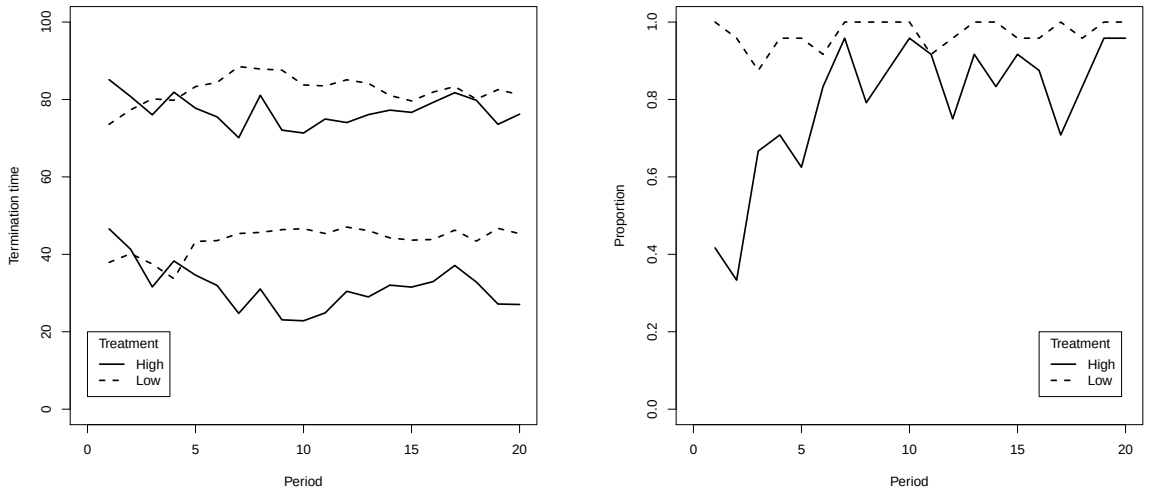
Table 1: Summary statistics of the participants in the experiment.

The lower part of this table shows the average intervals constructed and the average expected payment, where averages are taken over all individuals over all twenty periods and the expectation is based on the expected payment given the interval chosen on basis of the true distribution. The average interval in the low volatility treatment almost fully captures the interval that a risk neutral individual would optimally choose (when knowing the true distribution) and the mode of the true distribution. In the high volatility treatment a substantial

part of the risk neutral optimal interval is not captured in the average interval chosen; even the mode of the true distribution is just not contained. In both treatments subjects design longer intervals than a risk neutral individual would optimally do. The mis-positioning of the intervals in the high volatility treatment relative to the low volatility treatment, leads to subjects' expected payment being significantly higher in the low volatility treatment compared to the high volatility treatment (Mann-Whitney U: $p < 0.001$).

3.1 Choices

Panel (a) of Figure 3 presents the development of the average interval chosen during the course of the experiment for each of the two treatments. We see that there is some learning in the first periods and on average behavior stabilizes in the low volatility treatment while this is less so the case in the high volatility treatment. The earlier observed properties on the positioning of the intervals relative to the risk-neutral optimal intervals and the lengths of the intervals appears not to be an artefact of averaging over rounds but a persistent property. The risk-neutral optimal intervals have the property that the upper bound of the interval in the high volatility treatment should be equal to the lower bound of the interval in the low volatility treatment. Where averaged over time the former is 33.5 points above the latter, there is no time period in which these bounds differ by less than 24. The regression results presented in Table 2 indicate that over time the intervals marginally shrink in the low volatility treatment and marginally expand in the high volatility treatment. Furthermore, the choice of interval length does not correlate significantly with gender, risk attitude and cognitive ability.



(a) Average intervals over time.

(b) Share of intervals containing the true mode.

Figure 3: Average intervals over time and share of intervals containing the mode of the true distribution in the low volatility (dashed) and the high volatility (solid) treatment.

One property of the interval scoring rule is that if a subject's belief distribution over termination times is single-peaked, then the mode of this distribution should be contained in the reported interval. We see that for the low volatility treatment the mode of the true distribution (at 66) is during the whole course of the experiment contained in the average interval chosen; for the high volatility treatment, most of the time the mode (at 31) is not contained. Due to the flatness of the true distributions at the mode, it is hard for subjects to learn or to identify the true mode.[4] Allowing for a certain degree of mis-identification, panel (b) of Figure 3 shows the the share of intervals that contained the true mode at each time period. We classify each interval, that intersects with a termination time that is at least 95% as likely to realize as the true mode, as containing the true mode.[5] The figure shows that in the low volatility treatment in all periods at least 21 of the 24 subjects, and half of the time all 24 subjects, had an approximate mode contained in their interval. In the high volatility treatment more than half of the time more than 20 of the 24 subjects had an approximate mode contained in their interval, though in the first five periods less than 18 individuals qualify for this criterion. The fraction of subjects in the high volatility treatment that make a good forecast in this respect is never above this fraction in the low volatility treatment.

Relatedly, in Table 2 we study how individual characteristics relate to the interval lengths chosen as well as to the fact whether or not they contain the true mode. We see that none of the characteristics matter in the choice of interval length (first two columns). This is a surprising finding, given that we would expect risk attitudes to play a role in the choice of interval length. Moreover, we see that characteristics are not a significant predictor for the true mode being contained in the chosen interval (last two columns; coefficients are the marginal effects at the mean from a logit model).

All in all, individuals make better predictions (measured relative to the risk-neutral optimal interval and for the mode being contained in the interval) in the low volatility treatment compared to the high volatility treatment. There is no indication that this is due to any of the individual attributes. As the distinctive element of the treatments is the volatility, and as such the structure of the uncertainty, we can conclude that the nature of the uncertainty may have a large impact on individuals making good forecasts. Despite this sensitivity towards uncertainty, risk attitude seems not to be of importance – something we will get back to in Subsection 3.3. First, we explore how subjects adapt their chosen intervals on basis of experiences.

[4]Multiples of millions of simulations are needed to numerically identify the true mode. It is therefore not to be expected that our experimental subjects would be able to learn to do so within twenty rounds (even when taking into account that during one round they learn more about the process than one termination time).

[5]This implies that the range of values that could be considered as mode are [51,84] in the low volatility treatment and [25,40] in the high volatility treatment. Not allowing for mis-identification (i.e. only accepting the true mode), does not have any impact on the main findings.

	Interval length		Mode contained (marginal effect)	
Treatment	Low	High	Low	High
Constant	54.4004***	30.4989		
	(12.8491)	(18.2669)		
Period	−0.2736*	0.2930**	0.0021	0.0198***
	(0.1440)	(0.1166)	(0.0017)	(0.0034)
Gender	−2.0509	1.9541	0.0286	−0.0361
	(4.2102)	(4.7124)	(0.0248)	(0.0485)
Risk attitude	0.1538	0.4683	−0.0051	0.0090
	(1.0740)	(1.0332)	(0.0056)	(0.0171)
Cognitive ability	−0.3023	0.2065	−0.0013	0.0064
	(0.2521)	(0.3764)	(0.0013)	(0.0046)
Observations	480	480	480	480
R-squared	0.0446	0.0284		

Standard errors clustered on the individual level in parentheses
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 2: Interval length and whether the mode is contained in the interval against individual characteristics.

3.2 Dynamics of choices and learning

Each period, after having chosen their interval, subjects immediately experience the consequence of their choice. For our analysis on the dynamics of subjects' choices, which provides information on their learning, we distinguish four mutually exclusive and jointly exhaustive experiences, depending on the termination time of the time series relative to the chosen interval: (1) the termination time is below the interval, (2) the termination time is in the interval, (3) the termination time is above the interval, but the time series terminated before $t = 100$, and (4) the time series did not terminate before $t = 100$. We label these possible experiences by 'below', 'hit', 'above', and 'no hit', respectively (see Figure 4). Only the experience 'hit' yields a positive payoff; the other experiences do not yield any payoff.

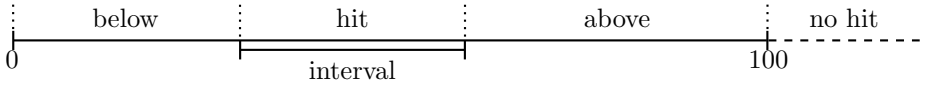


Figure 4: The four possible experiences.

We use the following regression model to estimate how individuals adapt their interval in response to their experiences:

$$\Delta b_{i,t} = \beta_0 + \beta_1 \, Below_{i,t-1} + \beta_2 \, Above_{i,t-1} + \beta_3 \, NoHit_{i,t-1} + \Gamma_i \, C_i + \varepsilon_{i,t}.$$

Here, $\Delta b_{i,t}$ denotes the change in either the upper or lower bound of the interval of individual i in period t. C_i is the vector that stores individual i's characteristics. Our main interest lies in the coefficients of β_1, β_2 and β_3 that capture the adjustment in the interval bound relative to

the 'hit' experience. The results are shown in Table 3 and indicate that subjects react quite significantly to previous period experiences.

	Low volatility		High volatility	
	lower bound	upper bound	lower bound	upper bound
Constant	5.0193^{**}	2.3811	3.2089	0.1320
	(2.4196)	(2.1233)	(2.3037)	(1.7443)
Below $(t-1)$	-15.4055^{***}	-8.2924^{*}	-11.4430^{***}	-4.7961^{***}
	(4.4908)	(4.1464)	(2.3877)	(1.6610)
Above $(t-1)$	2.5710	4.7751	0.5756	3.2335^{*}
	(5.3155)	(5.4829)	(1.8960)	(1.8387)
No hit $(t-1)$	-5.1643^{**}	-1.6478	-8.8576^{***}	-4.6007^{**}
	(1.8956)	(1.9694)	(2.5277)	(2.0867)
Gender	1.0421^{**}	0.6746	1.4763^{**}	0.0403
	(0.4715)	(0.4199)	(0.6159)	(0.4005)
Risk attitude	-0.0861	0.0218	-0.3772^{**}	-0.1077
	(0.1082)	(0.0958)	(0.1580)	(0.0864)
Cognitive ability	-0.0009	-0.0245	0.0595	0.0487
	(0.0309)	(0.0274)	(0.0449)	(0.0312)
Observations	456	456	456	456
R-squared	0.0548	0.0245	0.0922	0.0369

Standard errors clustered on the individual level in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 3: Interval updating depending on the experiences in the previous period.

When subjects experienced a termination below the selected interval in the previous period, in both treatments, they shift both bounds downwards, and the lower bound with a large amount compared to the upper bound, this leading to a widening of the interval. In case the series terminated above the chosen interval, individuals seem to shift both bounds upward in a manner that also yields a widening of the interval, but these effects are not statistically significant. The responses to these two experiences are consistent (for sure, not inconsistent) with Bayesian learning.

In the more extreme case where the time series did not terminate before $t = 100$ (the 'no hit' experience), in both treatments both interval bounds are shifted downwards (not significant for the upper bound in the low volatility treatment). This adjustment is clearly inconsistent with Bayesian learning. Individuals seem prone to the gambler's fallacy (cf. Lehrer, 2009) by acting in accordance to the mistaken belief that, in order to balance the mean, a no hit should be followed by an early hit. From the regressions reported in Table 8 in Appendix C it follows that most significant adjustments take place in the first half of the experiment which is in line with Bayesian learning (taking into account the difference in the amount of information to incorporate in the updating process); though, some adjustments persist in being significant.

From the adjustments in choices in response to experiences it is evident that subjects do learn throughout the experiment. Yet, one question is *what* they learn. In this section, we

more or less assumed that they try to learn to make better choices. Though, it may well be that they form better beliefs and that their choices are just a reflection of that. In Appendix B, under some additional structural assumptions on the subjects' beliefs, we investigate how beliefs are adjusting in response to experiences. Similar inferences are derived, though we find some differences in statistical significance of the reported effects.

3.3 Performance

In each treatment we measure individual performance in the prediction task as the ex ante expected payoff relative to the maximum ex ante expected payoff that can be obtained in the respective treatment. Here, the ex ante expected payoff refers to the expected payoff given the interval chosen and the incentives provided by the interval scoring rule and the true distribution over termination times as plotted in Figure 2; the maximum ex ante expected payoff is based on the same incentives provided and true distributions, but given that the risk neutral optimal interval is chosen. The columns labeled 'Normal' in Table 4 present the result of cross-sectional regressions of the individual performance on the participants' characteristics.

	Low volatility		High volatility	
	Normal	Conditional	Normal	Conditional
Constant	0.8355^{***}	0.9652^{***}	0.8057^{***}	0.8227^{***}
	(0.1464)	(0.0367)	(0.2151)	(0.0949)
2nd Half	0.0612^{***}	0.0192^{**}	0.0251	0.0320^{***}
	(0.0188)	(0.0068)	(0.0164)	(0.0106)
Gender	0.0437	0.0210	−0.0276	−0.0220
	(0.0349)	(0.0158)	(0.0421)	(0.0191)
Risk attitude	-0.0191^{**}	-0.0060^{**}	−0.0150	0.0055
	(0.0082)	(0.0029)	(0.0121)	(0.0054)
Cognitive ability	0.0014	0.0002	−0.0011	0.0005
	(0.0028)	(0.0007)	(0.0050)	(0.0017)
Observations	480	480	480	480
R-squared	0.0870	0.0415	0.0517	0.0289

Standard errors clustered on the individual level in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 4: Individual prediction performance with performance measured as ex ante expected payoff relative to maximum possible payoff (normal) and the latter maximum conditional on chose interval length (conditional).

The only individual characteristic that appears to have a significant impact on performance is risk attitude, and this effect is only significant in the low volatility treatment. The negative sign indicates that more risk averse individuals make better predictions. Gender and cognitive ability are not a significant predictor for individual performance.

In principle the subjects' interval choices can be disentangled in two choices: the length of the interval and its location. In order to disentangle the impact of individual characteristics

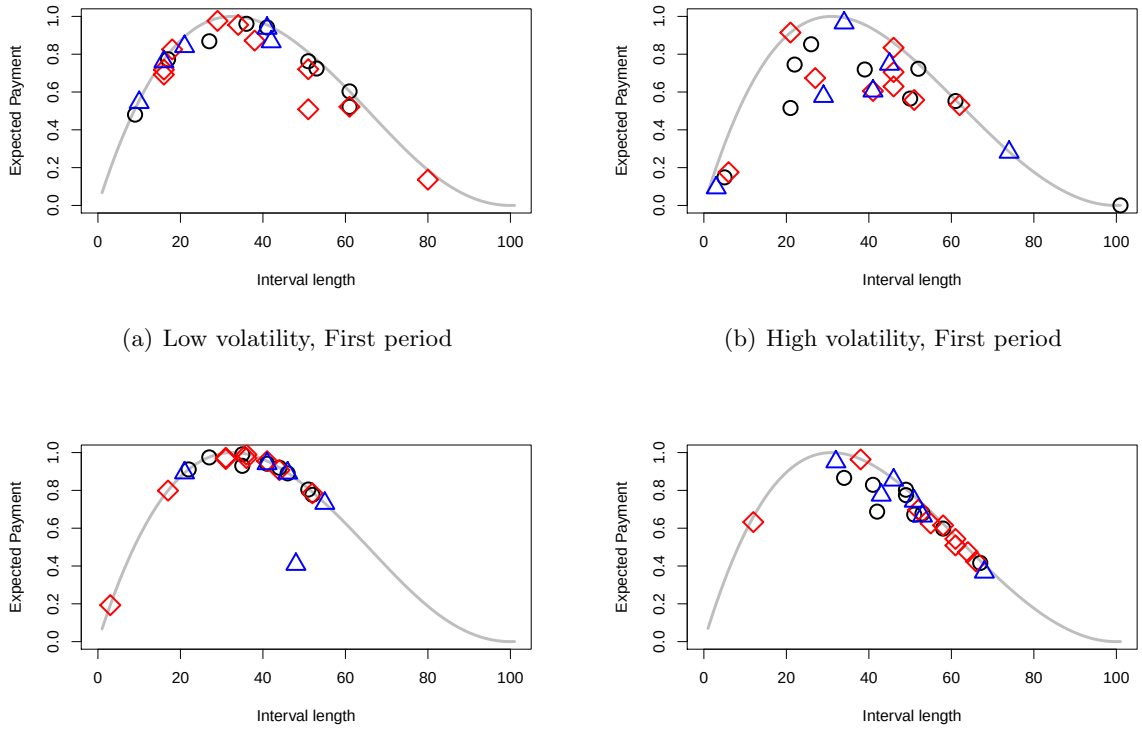
on performance (or the lack thereof) for these two choices, we ran additional regressions where the performance is measured relative to the maximum possible payoff given the chosen length of the interval, the result of which are presented in the column label 'Conditional'. Again, the only individual characteristic that has a significant impact on performance is risk attitude, and, again, this effect is only significant in the low volatility treatment. This indicates that part of the effect of risk attitude on performance can be attributed to the location of the interval. As the impact of risk attitude on interval length was shown to be highly insignificant (recall Table 2), we may conclude that the worse performance of risk seeking individuals can be fully attributed on where they locate their intervals.

Comparing the variables 'Constant' and '2nd Half' across treatments, the coefficients suggest that individuals perform better in the low volatility treatment compared to the high volatility treatment and that performance has improved during the experiment in both treatments. This improvement is not significant in the high volatility treatment, but is so when the performance is measured relative to the chosen interval lengths. This suggests that in the high volatility treatment subjects mainly improve in their choice of interval location.

In order to draw a better picture of how risk attitude affects performance and how this interacts with the volatility of the stochastic process, Figure 5 displays individual performance (y-axis) conditional on interval length (x-axis) for the low and high volatility treatments in the first and last period of decision making. Panels (a) and (b) show first period choices for the low and high volatility treatments respectively, while panels (c) and (d) show the same individuals' choices in the last period. The curves in the plots identify the (normalized) maximum attainable payoff as a function of chose interval length. Three different geometric are used to distinguish individuals from three different risk groups where, for each treatment, we split the subjects at the one-thirds and two-thirds quantile of their reported scores. In the figure, the circles refer to the individuals with the lowest risk tolerance, the diamonds to those with medium risk tolerance, and the triangles to those with the highest risk tolerance.

Comparing the performances in the first and last period, we see that the figure nicely illustrates the effects observed in Table 4. In the low volatility treatment, with the geometric shapes being close to the curves in panel (a) and (c), subjects succeed to choose the location close to optimal given the chosen interval length already in the first period and still do so in the last period. Though, comparing the distribution of interval lengths over these two panels, we see that over time subjects improve in their choice of interval length (while they keep choosing the right location given the length). Moreover, there is no apparent difference in the distribution of interval lengths across risk groups (which we saw already in Table 2).

In the high volatility treatment, we do not observe the same effect (panel (b) and (d)). First, subjects do not succeed to choose the best location given the chosen interval length in the first period, but learn to do so over time. Second, while like in the low volatility



(a) Low volatility, First period

(b) High volatility, First period

(c) Low volatility, Last period

(d) High volatility, Last period

Figure 5: Individual performance against interval length for the two treatments in the first and last period.

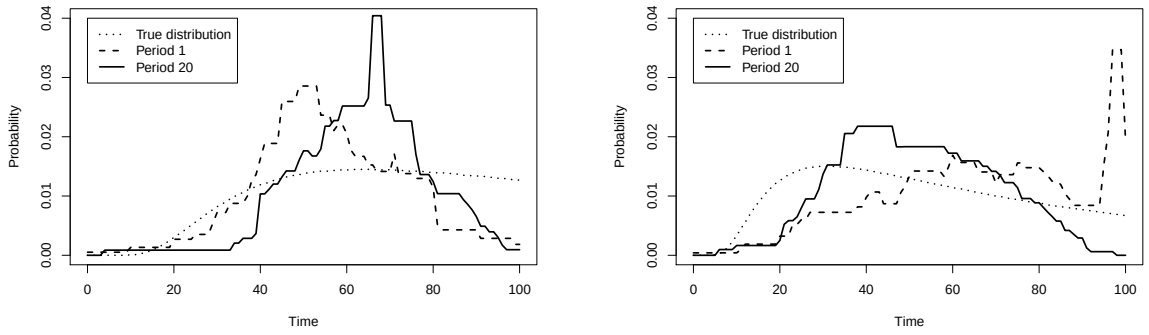
treatment the dispersion of interval lengths is reduced over time, we see that they cluster on a suboptimal level: subjects opt for too lengthy intervals. Overall, this explains the lack of improvement in individual performance over time in this treatment. Again, there is no apparent difference in the distribution of interval lengths across risk groups.

3.4 Aggregate forecasts of the underlying distribution

Even though the time series shown to the participants are identical and they thus only possess common information about the underlying parameters we observe significant variation in the forecasted intervals. The aggregation of interval predictions of several subjects yields a distribution over possible termination times. Such an amalgamation of individual forecasts may provide a better forecast than any of the individual forecasts.

Figure 6 shows the aggregated probability density functions for the two treatments in the first (dashed line) and last period (solid line) of the experiment, where for each treatment the aggregation is taken over all 24 participating subjects. Table 5 shows some key summary statistics related to these densities. Overall, we see that the aggregate forecasts improve over

time. Next, we focus on the quality of an aggregated prediction in relation to group size, and how the quality develops over time.



(a) Low volatility treatment.

(b) High volatility treatment.

Figure 6: Distributions over termination times conditional on termination before time $t = 100$: true distribution (dotted) and aggregated interval choices in first (dashed) and last (solid) round of experiment.

	Low volatility			High volatility		
	First round	Last round	True distr.	First round	Last round	True distr.
Mean	55.75	63.29	63.58	65.85	51.62	51.13
Median	54.73	65.11	64.64	67.14	51.03	49.02
Std. dev.	16.96	15.05	21.54	23.01	17.47	24.12

Table 5: Key summary statistics related to the distributions in Figure 6.

In order to study the impact of group size (and composition) on the quality of predictions when aggregating individual predictions over groups it is important to adopt a good measure to quantify 'quality of prediction'. One property that such a measure should capture is that it allows for a fair comparison within and across groups of different sizes. In our analysis, we will make use of the *Hellinger distance* (Hellinger, 1909) that quantifies the similarity between two probability distributions. An important advantage of the Hellinger distance over often used alternatives (such as the Kullbeck-Leibler divergence) is that it does not require absolute continuity, a property that is violated almost by design.[6]

The Hellinger distance of the (discrete) empirical probability distribution $Q = (q_1, \ldots, q_m)$ to the (discrete) true probability distribution $P = (p_1, \ldots, p_m)$ is defined as

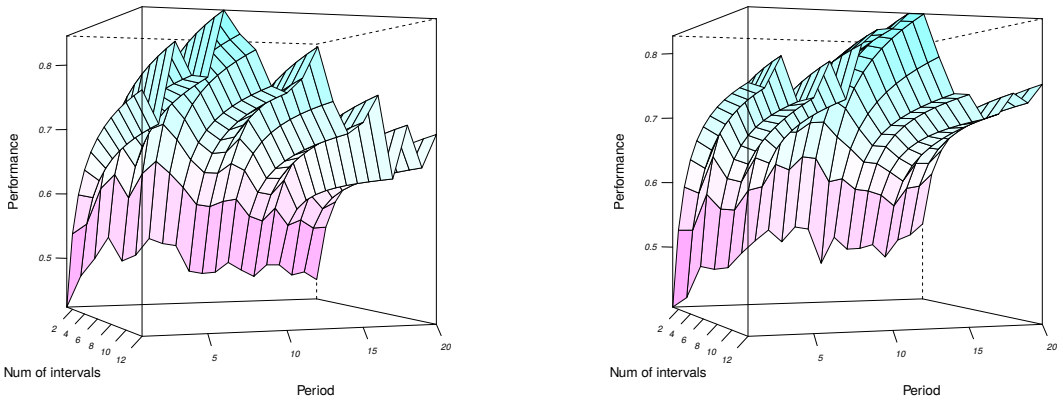
$$H(Q, P) = \frac{1}{\sqrt{2}} \sqrt{\sum_{j=1}^{m} (\sqrt{q_j} - \sqrt{p_j})^2}.$$

[6]Yet another desirable property of the Hellinger distance, that we do not exploit here, is that it satisfies the triangular inequality.

In case the two distributions P and Q coincide, the Hellinger distance equals zero. The maximum Hellinger distance of one is obtained when the supports of the two distributions are disjoint. Consequently, for intuitive reasons, we henceforth define a performance index, Z, that equals one minus the Hellinger distance:

$$Z(Q, P) = 1 - H(Q, P).$$

In Figure 7 we plot the performance measure, Z, of the aggregated interval predictions over different group sizes and time periods. In the three dimensional graph, each point represents the average performance for a given aggregation size (increasing from far to near) and time period (increasing from left to right). The left panel shows this for the low volatility treatment and the right panel for the high volatility treatment. The graphs from both treatments look quite similar and it is evident that the performance improves substantially when increasing the group size. In both treatments, for given group size, the performance averaged over all possible groups of that size is rather constant over time. Any effects of learning that we saw to be present on the individual level, in particular during the first eight periods, seem to have disappeared in the aggregation process.



(a) Low volatility treatment

(b) High volatility treatment

Figure 7: Average performance of aggregate interval predictions over group size and over time.

In the following, we quantify the impact of groups size on forecasting performance using a regression model. For each possible coalition of individuals of group size between two and twelve, we compute the coalition's forecasting performance in each period and store the coalition's average values for gender, risk attitude and cognitive ability. This yields for each treatment a dataset with almost 200 million entries.[7] Next, we regress performance on group

[7]There are 20 time periods with 24 individuals in each treatment. All possible group configurations of individuals in group sizes between one and twelve equal 9,740,685.

size, using as regressors either the individual characteristics (model (1)) or individual dummies (model (2)). Table 6 presents the results of these regressions.

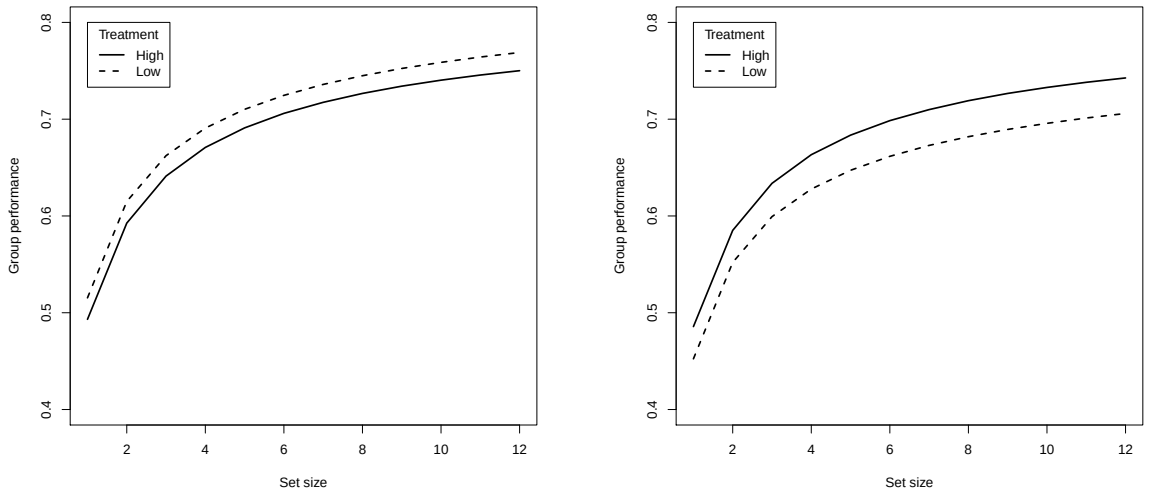
	Low volatility		High volatility	
	(1)	(2)	(1)	(2)
Constant	0.6158^{***}		0.3296^{***}	
	(0.0030)		(0.0027)	
2nd Half	-0.0630^{***}	-0.0630^{***}	-0.0075^{***}	-0.0075^{***}
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Gender	-0.0357^{***}		-0.0081^{***}	
	(0.0000)		(0.0000)	
Risk attitude	0.0066^{***}		0.0182^{***}	
	(0.0000)		(0.0000)	
Cognitive ability	-0.0029^{***}		0.0014^{***}	
	(0.0000)		(0.0000)	
Group size 2	0.0996^{***}	-0.4164^{***}	0.0994^{***}	-0.3941^{***}
	(0.0031)	(0.0058)	(0.0029)	(0.0053)
Group size 3	0.1469^{***}	-0.8850^{***}	0.1480^{***}	-0.8391^{***}
	(0.0030)	(0.0086)	(0.0028)	(0.0078)
Group size 4	0.1754^{***}	-1.3726^{***}	0.1776^{***}	-1.3030^{***}
	(0.0030)	(0.0115)	(0.0027)	(0.0104)
Group size 5	0.1948^{***}	-1.8691^{***}	0.1978^{***}	-1.7763^{***}
	(0.0030)	(0.0144)	(0.0027)	(0.0130)
Group size 6	0.2092^{***}	-2.3706^{***}	0.2128^{***}	-2.2549^{***}
	(0.0030)	(0.0172)	(0.0027)	(0.0156)
Group size 7	0.2205^{***}	-2.8754^{***}	0.2242^{***}	-2.7370^{***}
	(0.0030)	(0.0201)	(0.0027)	(0.0182)
Group size 8	0.2295^{***}	-3.3823^{***}	0.2334^{***}	-3.2213^{***}
	(0.0030)	(0.0230)	(0.0027)	(0.0208)
Group size 9	0.2370^{***}	-3.8908^{***}	0.2409^{***}	-3.7074^{***}
	(0.0030)	(0.0258)	(0.0027)	(0.0234)
Group size 10	0.2433^{***}	-4.4004^{***}	0.2471^{***}	-4.1947^{***}
	(0.0030)	(0.0287)	(0.0027)	(0.0261)
Group size 11	0.2488^{***}	-4.9110^{***}	0.2524^{***}	-4.6830^{***}
	(0.0030)	(0.0316)	(0.0027)	(0.0287)
Group size 12	0.2535^{***}	-5.4222^{***}	0.2569^{***}	-5.1720^{***}
	(0.0030)	(0.0345)	(0.0027)	(0.0313)
Individual effects	No	Yes	No	Yes
Avg. ind. effect		0.5160		0.4935
Observations	194,813,700	194,813,700	194,813,700	194,813,700
R-squared	0.2119	0.2757	0.0502	0.1389

Standard errors in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 6: Regression analysis of group performance on group size.

The effect of groups size is for both treatments similar in both model specifications. To see this, first, notice that the values of the individual attributes (multiplied with the average values presented in Table 1) added to the constant in models (1), approximately equals the average individual effect presented in models (2); and, second, notice that the coefficients for 'Group size k' in model (1) are close to the coefficients of the same variable in model (2)

after adding $k-1$ times the average individual effect. Figure 8 presents the performance as a function of groups size for the two treatments during the first end second half of the experiment as they can be retrieved from the coefficients estimated in the regression.



(a) Group performance, First half

(b) Group performance, Second half

Figure 8: Group performance.

In both treatments we find that performance is increasing in group size at a decreasing rate. Throughout the first half of the experiment the performance is (for all group sizes) about 0.020 higher in the low volatility treatment while in the second half performance is (for all group sizes) about 0.035 higher in the high volatility treatment. We believe that in the first half the low volatility treatment benefits from individuals making better forecasts, while in the second half the high volatility treatment benefits from more variation in individual choices which may produce more structure in the aggregate distributions.

4 Conclusion

In this paper we propose and implement a non-market based mechanism to elicit forecasts and to test it experimentally. In our experiment subjects have, incentivized by means of the interval scoring rule, to forecast, over a sequence of twenty periods, the termination time of a time series that is to be generated from a fixed but unknown random process by specifying an interval where they believe the time series is going to terminate. We study the choices individuals make in this environment, how these choices change over time in response to recent experiences, how individual forecasting performance relates to the level of the underlying uncertainty and individual attributes like cognitive ability, risk and gender, and the quality of the performance of group predictions depending on the size and composition of the group.

We find that individuals make better predictions in the low volatility treatment compared to the high volatility treatment, and there is no indication that this is due to any of the individual attributes. Over time individuals improve their forecasting performance in the low volatility treatment, but fail to do so significantly in the high volatility treatment where they learn to improve in the choice of location of the interval given the interval length, but fail to choose the correct length. Although they seem to learn by experience in a way consistent with Bayesian learning, subjects feel prone to the gambler's fallacy. All in all, on basis of individual choices, we can conclude that the nature of the uncertainty has a large impact on individuals' forecasts.

When aggregating forecasts over groups of individuals, we find that the group performance is increasing in group size at a decreasing rate. While, for any given group size, group performance is better in the low volatility treatment throughout the first half of the experiment, aggregated forecasts are better in the high volatility treatment during the second half. This is possibly due to there being less correlation in individual forecasting errors, which makes the aggregate forecast resemble the underlying distribution better. Although we believe we may conclude that the mechanism studied yields a quite good forecast already when aggregating over few individuals, and provides evidence in favor of the use of non-market based forecasting mechanisms, one systematic concern of this method is the underestimation of the tails of the distribution – an issue that asks for further innovations in the design of elicitation methods.

References

1. Arrow, K. J., R. Forsythe, M. Gorham, R. Hahn, R. Hanson, J. O. Ledyard, et al. (2008). The promise of prediction markets. Science 320(5878): 877-878.

2. Brier, G. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review 78: 1-3.

3. Budescu, D. V. and E. Chen (2014). Identifying expertise to extract the wisdom of the crowds. Management Science, forthcoming.

4. Cowgill, B. and E. Zitzewitz (2013). Corporate prediction markets: Evidence from Google, Ford and Firm X. Working paper.

5. Dohmen, T., A. Falk, D. Huffman and U. Sunde (2010). Are risk aversion and impatience related to cognitive ability? The American Economic Review 100(3): 1238-1260.

6. Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp and G. Wagner (2011). Individual risk attitudes: Measurement, determinants and behavioral consequences. Journal of the European Economic Association 9(3): 522-550.

7. Fischbacher, U. (2007). zTree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10(2): 171-178.

8. Forsythe, R., F. Nelson, G. R. Neumann and R. Wright (1992). Anatomy of an experimental political stock market. American Economic Review 82(5): 1142-1161.

9. Galbiati, R., K. Schlag, and J. van der Weele (2013). Sanctions that signal: An experiment. Journal of Economic Behavior and Organization 94: 34-51.

10. Gillen, B. J., C. R. Plott and M. Shum (2013). Inside Intel: Sales forecasting using an information aggregation mechanism. Working paper.

11. Goel, S., D. M. Reeves, D. J. Watts and D. M. Pennock (2010). Prediction without markets. Proceedings of the ACM EC'10 Conference, Cambridge, MA.

12. Goldstein, D. G., R. P. McAfee and S. Suri (2014). The wisdom of smaller smarter crowds. Proceedings of the ACM EC'14 Conference, Stanford, CA.

13. Greiner, B. (2004). An online recruitment system for economic experiments. In: K. Kremer and V. Macho (eds.): Forschung und wissenschaftliches Rechnen 2003. GWDG Bericht 63, Göttingen: Ges. für Wiss. Datenverarbeitung: 79-93.

14. Hanson, R., R. Oprea and D. Porter (2006). Information aggregation and manipulation in an experimental market. Journal of Economic Behavior and Organization. 60: 449-459.

15. Harrison, G. W., J. Martinez-Correa, T. Swarthout and E. R. Ulm (2013). Scoring rules for subjective probability distributions.

16. Hellinger, E. (1909). Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. Journal für die reine und angewandte Mathematik 136: 210-271.

17. Hopman, J. W. (2007). Using forecasting markets to manage demand risk. Intel Technology Journal 11(2): 127-135.

18. Lehrer, J. (2009). How We Decide. New York: Houghton Mifflin Harcourt.

19. Peeters, R., M. Vorsatz and M. Walzl (2012). Beliefs and truth-telling: A laboratory experiment. Working paper.

20. Schlag, K. and J. van der Weele (2009). Efficient interval scoring rules. Working paper.

21. Smith, V. (1962). An experimental study of competitive market behavior. Journal of Political Economy 70(2): 111-137.

22. Tausch, F., J. Potters and A. Riedl (2014). An experimental investigation of risk sharing and adverse selection. Journal of Risk and Uncertainty 48: 167-186.

23. Veiga, H. and M. Vorsatz (2006). Price manipulation in an experimental asset market. European Economic Review 53: 327-342.

24. Winkler, R. and A. H. Murphy (1970). Nonlinear utility and the probability score. Journal of Applied Meteorology 9(1): 143-148.

A Experimental instructions

Welcome

You are about to participate in a session on individual decision-making. Thank you for agreeing to take part. The session should last 60 to 90 minutes.

You should already have turned off all your mobile phones, smart phones, mp3 players and any such devices. If not, please do so immediately. These devices must remain switched off throughout the session. Place them in your bag or on the floor besides you. Do not have them in your pocket or on the table in front of you.

The entire session will take place through the computer. You are not allowed to talk or to communicate with other participants in any other way during the session.

You are asked to abide by these rules throughout the session. Should you fail to do so, we will have to exclude you from this (and future) session(s) and you will not receive any compensation for this session.

We will start with a brief instruction period. Please read these instructions carefully. They are identical for all participants in this session with whom you will interact. If you have any questions about these instructions or at any other time during the experiment, then please raise your hand. One of the experimenters will come to answer your question.

Compensation for participation in this session

In addition to the 3.00 Euro participation fee, what you will earn from this session will depend on your decisions and chance. In the instructions and all decision tasks that follow, payoffs are reported in Experimental Currency Units (ECUs). At the end of the experiment, the total amount you have earned will be converted into Euros using the following conversion rate:

1 ECU = 6 Eurocents.

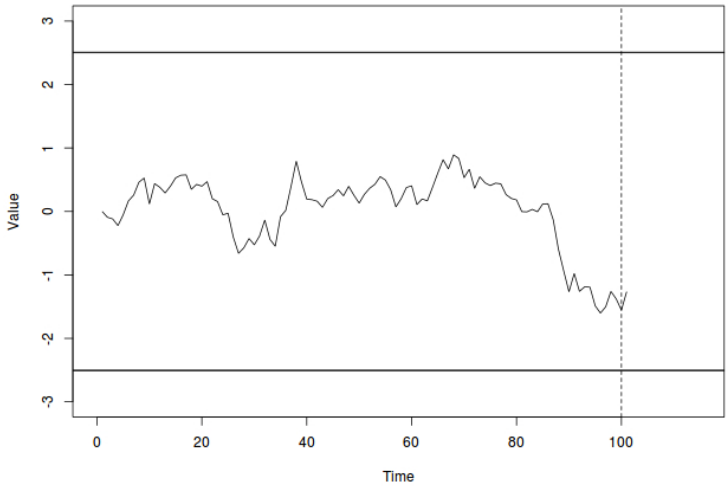
The payment takes place in cash at the end of the experiment. Your decisions in the experiment will remain anonymous.

Instructions

This session consists of twenty rounds. Each round you are faced with two decision tasks and the payoff (in ECU) that you collect depends on the decisions you make and chance. At the end of the session you are paid according to eight random draws (with replacement) over the payoffs you earned over the two tasks in the twenty rounds.[8]

[8]To elaborate, in total you make 40 decisions that lead to 40 payoffs. From these 40 payoffs, eight are drawn for actual payment. These draws are taken with replacement, meaning that it is not excluded that the same payoff is drawn multiple times, and for each participant individually.

Before the first round starts, you will be shown the *time series* that results from some *random process*. See the figure below for an example of such a time series.



The random process from which the time series has been generated is kept fixed during the entire session, but every round a different time series will be generated using the same random process. Each round you will see a new time series; so, you will get better acquainted with the random process over rounds. Apart from the realized time series in the previous rounds and the time series shown to you at the beginning (and the one in the figure above), no further information will be given, except that the time series will start at a value of 0 at time $t = 0$.

Each round, before you see the time series that is generated for that round, you are faced with two prediction tasks:

1. First, you are asked how likely you regard the event that the time series hits the *boundary* (one of the thick horizontal lines in the figure above) before time $t = 100$. You can express your expectation regarding this event by moving the triangular cursor along the line. See the figure below.

   The payoff that you earn with this decision task depends on the point you select along the line and the generated time series. The potential payoffs in the event that the time series hits the boundary before time $t = 100$ and in the event that it does not are shown on-screen in real-time when you move the cursor along the line.

2. Second, conditionally on the time series hitting the boundary before time $t = 100$, you are asked to indicate within which *time interval* you think the time series will hit

the boundary. You can indicate this interval by moving two triangular cursors (one indicating the lower bound of the interval; the other indicating the upper bound of the interval) along the time line. See the figure below.

Only in the event that the time series hits the boundary within the indicated time interval you collect a payoff. The smaller the interval that you indicated, the larger this potential payoff is. This potential payoff is shown on-screen in real-time when you move the cursors along the time line.

3. To avoid the unfortunate event that you confirm your decisions while not being completely confident these being the right decisions, you have to approve your decisions at the bottom of the screen.

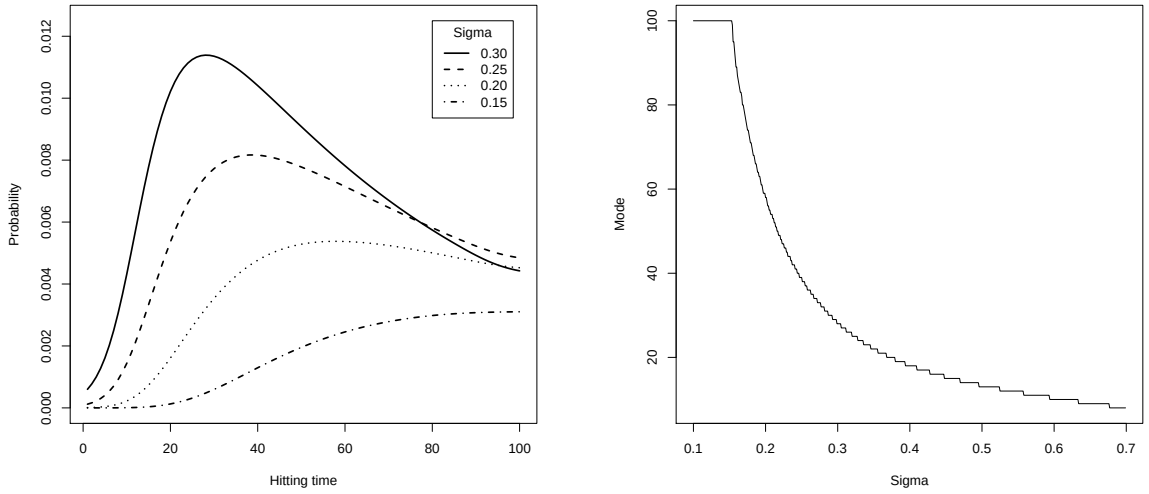
After having made your predictions, the time series generated for that round will be shown to you. Furthermore, you are informed about the payoffs you collected. It is important to note here that the time series is generated by a statistical software package and is not manipulated for the purpose of this experiment. As all time series shown to you are generated from the same random process, over rounds you will gradually become more familiar with the underlying process.

B Dynamics of beliefs and learning

In Subsection 3.2 we investigated how individual adjust their choices in response to recent experiences. The reason that adjustments are expected is that subjects learn. One question is *what* they learn. In Subsection 3.2 we more or less assumed that they learn to make better choices. Though, it may well be that they form better beliefs and their choices are just a reflection of that.

In this appendix we investigate how they adjust their beliefs. To do so, we assume that subjects know that the time series involves a gradual increment of random draws from a a normal distribution with zero mean and variance of some fixed σ. Next, we study their adjustments in beliefs on σ. In order to be able to do so, we have to map their interval choices to intervals of beliefs that are consistent with the chosen intervals. For this we use the property of the interval scoring rule that in case a subject has a single-peaked belief distribution, the stated interval should contain the mode of this distribution.

Panel (a) of Figure 9 shows the distribution over termination times for different values of σ. We see that the distribution shifts leftwards if σ increases, and so does the mode of the distribution. Panel (b) plots the relation between σ and the mode of the distribution. We use this relation to map chosen intervals to intervals of beliefs over σ. For example, the σ-s that are compatible with the interval $[20, 60]$, are precisely the σ-s in the interval $[0.196, 0.379]$. After all, only σ-s in this interval produce a distribution of which the mode is in $[20, 60]$.



(a) Full distribution, few volatility levels.

(b) Many volatility levels, modes of distribution.

Figure 9: Relationship between volatility of process (sigma) and the implied distribution of conditional termination times.

Table 7 shows a similar regression as presented in Table 3, but now on belief intervals

rather than choice intervals.[9] The findings are quite similar. Consistently with the findings reported earlier, subjects tend to adjust their intervals in conformation with Bayesian learning when the time series terminated below or above the stated interval and are prone to the gambler's fallacy when the time series did not terminate.[10] The main difference are that some effects that were significant are no longer significant while some effects that were not significant before turn out to be significant now. These changes in significance may be due to the nonlinear relation between the σ-s and the mode.

	Low volatility		High volatility	
	lower bound	upper bound	lower bound	upper bound
Constant	-0.0460^{*}	-0.0045	-0.0498	-0.0025
	(0.0231)	(0.0043)	(0.0659)	(0.0039)
Below $(t-1)$	0.1212^{**}	0.0061	0.1333^{***}	0.0126^{***}
	(0.0541)	(0.0077)	(0.0467)	(0.0043)
Above $(t-1)$	-0.0670	-0.0385	0.0023	-0.0101^{**}
	(0.1711)	(0.0403)	(0.0779)	(0.0046)
No hit $(t-1)$	0.0858^{*}	0.0023	0.2117^{**}	0.0097^{**}
	(0.0494)	(0.0040)	(0.0915)	(0.0038)
Gender	-0.0192	-0.0012	-0.0368	0.0001
	(0.0114)	(0.0009)	(0.0218)	(0.0010)
Risk attitude	0.0011	-0.0001	0.0082^{*}	0.0001
	(0.0016)	(0.0002)	(0.0046)	(0.0002)
Cognitive ability	-0.0007	0.0001	-0.0020	-0.0000
	(0.0008)	(0.0001)	(0.0015)	(0.0001)
Observations	456	456	456	456
R-squared	0.0292	0.0227	0.0391	0.0477

Standard errors clustered on the individual level in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 7: Belief updating depending on the realization in the previous period.

[9] A similar regression with a time dummy to assess the persistence of adjustment behavior is presented in Table 8.

[10] Realize here that the lower (upper) bound on choices map to the upper (lower) bound on beliefs.

C Additional table

	Choices				Beliefs			
	Low volatility		High volatility		Low volatility		High volatility	
	lower bound	upper bound	lower bound	upper bound	lower bound	upper bound	lower bound	upper bound
Constant	7.8115**	3.1941	3.1223	0.4234	−0.1174*	−0.0019	−0.0874	−0.0021
	(2.9770)	(2.2430)	(2.1918)	(1.6619)	(0.0645)	(0.0053)	(0.0815)	(0.0038)
Below $(t-1)$ $[\beta_1]$	−34.8335***	−9.4286	−11.1043***	−5.1415**	0.2238**	0.0057	0.1687**	0.0145**
	(6.6324)	(11.7972)	(2.6467)	(1.9835)	(0.1044)	(0.0344)	(0.0721)	(0.0058)
Above $(t-1)$ $[\beta_2]$	−0.7875	3.8952	8.7019*	16.3705***	0.0103	−0.0412	−0.5780	−0.0625***
	(5.2058)	(5.1633)	(4.7708)	(4.8814)	(0.1932)	(0.0398)	(0.3521)	(0.0104)
No hit $(t-1)$ $[\beta_3]$	−8.8623***	−2.3637	−12.0994***	−8.1480**	0.1720	0.0010	0.3637**	0.0160**
	(2.8655)	(2.6477)	(3.6091)	(3.3221)	(0.1019)	(0.0074)	(0.1649)	(0.0070)
2nd Half	−6.3241*	−1.6646	0.1780	−0.7232	0.1463	−0.0051	0.0846	−0.0001
	(3.4120)	(3.1324)	(2.0010)	(2.0045)	(0.0996)	(0.0058)	(0.0604)	(0.0034)
Below $(t-1)$ × 2nd Half $[\beta_4]$	25.1048***	1.9692	−0.8509	0.6754	−0.0300	−0.0027	0.0163	−0.0048
	(7.4402)	(12.2508)	(5.1225)	(3.4256)	(0.0283)	(0.0349)	(0.0214)	(0.0069)
Above $(t-1)$ × 2nd Half $[\beta_5]$			−9.4627**	−14.8506***			0.7174*	0.0605***
			(4.5297)	(5.0775)			(0.3515)	(0.0097)
No hit $(t-1)$ × 2nd Half $[\beta_6]$	7.0057*	1.3633	6.4748*	7.1121*	−0.0172	−0.0026	−0.2218**	−0.0126**
	(4.0806)	(3.7588)	(3.2602)	(3.9917)	(0.0146)	(0.0048)	(0.0960)	(0.0060)
Gender	0.9328**	0.6531	1.5075**	0.0996	−0.0173	−0.0013	−0.0398*	−0.0001
	(0.4309)	(0.4127)	(0.5591)	(0.4338)	(0.0103)	(0.0009)	(0.0204)	(0.0015)
Risk attitude	−0.0587	0.0250	−0.4013**	−0.1528	0.0009	−0.0001	0.0104**	0.0003
	(0.0989)	(0.0926)	(0.1496)	(0.0953)	(0.0017)	(0.0002)	(0.0048)	(0.0004)
Cognitive ability	0.0099	−0.0231	0.0630	0.0565*	−0.0008	0.0001	−0.0024*	−0.0001
	(0.0252)	(0.0267)	(0.0382)	(0.0280)	(0.0008)	(0.0001)	(0.0014)	(0.0001)
F-test (p-values)								
$H_3 : (\beta_1 + \beta_4) = 0$	0.0439	0.0948	0.0122	0.1350	0.0755	0.7254	0.0242	0.1401
$H_4 : (\beta_2 + \beta_5) = 0$			0.7214	0.5271			0.1594	0.5885
$H_5 : (\beta_3 + \beta_6) = 0$	0.4881	0.7229	0.0212	0.6692	0.1366	0.7278	0.0640	0.2828
Observations	456	456	456	456	456	456	456	456
R^2	0.0723	0.0251	0.1099	0.0669	0.0413	0.0235	0.0954	0.1089

Standard errors clustered on the individual level in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 8: Interval updating depending on the experiences in the previous period. First four columns refer to choice intervals, last four columns to belief intervals.