Can instant feedback attenuate Overconfidence?

UNIVERSITY OF GOTHENBURG
SCHOOL OF BUSINESS, ECONOMICS AND LAW

Author:
Babak Hakimi
gushakba@student.gu.se

Supervisor:
Dawei Fang, Ph.D

Graduate School
University of Gothenburg
School of Business, Economics and Law
Sweden
Spring 2019

Abstract

Overconfidence is one the most common behavioral biases. In this paper I make use of an online survey experiment to investigate whether overconfidence can be attenuated by providing individuals with feedback. This is done in two stages. In the first stage, a group of subjects receive feedback on how they perform relative to their estimated performance in terms of calibration. In the second stage, subjects are asked to provide interval estimates with 90% certainty to interval judgments. Between the first and second stage feedback is given to subjects in the treatment group in order to find out if they behave differently. The results from the experiment provide weak evidence in support of feedback attenuating overconfidence.

Keywords: *Behavioral Economics, Overconfidence, Decision-making*

Acknowledgements

I would like to thank my supervisor, Professor Fang, for guiding me through this work. His help has enabled me to develop the process of carrying out the experiment and also the work following that. A special thanks to Kristofer Heinl for helping me create the survey with the Qualtrics software and brainstorm on the contents of it.

Contents

1. Introduction

Overconfidence is the tendency for people to overestimate their knowledge, abilities, and the precision of their information, or to be overly sanguine of the future and their ability to control it (Ackert and Deaves 2010). A study found that 93% of American, and 69% of Swedish drivers, rate themselves as above-average drivers (Svenson 1981). The empirical literature suggests that overconfidence can cause individuals to make suboptimal financial decisions such as excessive trading, taking excessive risk, analyst exhibiting excessive optimism, and that men are more overconfident than women (Barber and Odean 2000; Glaser and Weber 2007; Jegadeesh and Kim 2006; Ackert and Deaves 2010).

In this study I seek out to investigate the possibility of learning from past experiences, and mistakes, in order to attenuate overconfidence. The question asked is *can instant feedback attenuate overconfidence in a subject's decision-making?* In order to investigate this research question an experiment is set up to find out if there is an effect to be measured from feedback. The experiment is designed in a survey-based fashion. In the first part, overconfidence is measured in terms of calibration by subjects estimating their performance relative to actual performance. A random generator either displays this calibration to subjects or not, dividing the sample into a treatment group and a control group. In the second part, subjects are asked to answer interval estimate questions with a confidence level of 90%. Here overconfidence is measured by the spread of the interval, if it is considered too narrow given 90% certainty. I analyze whether subjects that exhibit overconfidence in the calibration part, and receive feedback, show less overconfidence in the interval estimates part compared to their peers. The results from the experiment give weak evidence to support the hypothesis that feedback can attenuate overconfidence. There are signs of feedback having effects in both directions for over- and underconfidence, causing the intent of feedback to be undermined.

1.1 Research Purpose

The motivation behind this research is to look at ways to mitigate effects of overconfidence in decision-making. The literature on overconfidence is widely accepted as an established part of behavioral science. The previously cited effects can be observed from different experiments from research within the field of decision-making. It is interesting to see whether feedback can play a part in remedying the previously observed effects. The goal is to investigate whether overconfidence can be self-altered and improved by providing information on it. Such an effect can have immense benefits to individuals in the decision-making process. The approach in this study opens up for an avenue to see if this could be the case. A simple feedback mechanism, as used in this study, is an easy implementation in different areas of decision-making. With today's technology such an element could be integrated into software programs, and devices of different kinds, in order to improve the decision-making process for individuals. As earlier research indicates, the largest area in need of improvement of overconfidence is in interval judgments. This study casts a light on two aspects of overconfidence. On one hand, to see whether overconfidence is transferable from calibration to interval judgments. On the other hand, if so, can feedback on calibration help improve the judgments of interval estimates?

2. Literature Review

An early study from Lichtenstein and Fischhoff (1977) find that people's probability judgments are prone to systemic bias. The most common bias they find is overconfidence. The aim of the study is to investigate whether the amount of knowledge a subject possess effect their calibration. In one experiment (experiment 3; Lichtenstein and Fischhoff 1977) they compare groups with different levels of knowledge based on the number of correct answers. Subjects are divided into three groups, best, middle, and worst. Although all groups tend to be overconfident, Lichtenstein and Fischhoff (1977) find that the more you know the better your calibration. They also find evidence for people's calibration differing with varying degrees of difficulty with respect to item tasks. A common finding across experiments is that when items are divided into easy/hard, subjects demonstrate underconfidence for the easy items and overconfidence for the hard ones (Lichtenstein and Fischhoff 1977).

In their paper, *Overconfidence in Interval Estimates*, Klayman and Soll (2003) give a brief history of overconfidence where they elaborate on the so-called hard-easy effect. Binary choice tasks show a hard-easy effect where overconfidence attenuates, and sometimes reverses, as percentage correct increases (Klayman and Soll 2003). Concurrently, a different picture emerges on confidence in a binary choice setting. A new view suggests that judge's responses contain error but with little to no bias (Klayman and Soll 2003). In an earlier paper Klayman et al (1999) find that binary choice questions elicit a modest bias toward overconfidence whilst subjective confidence intervals elicit large bias. Furthermore, Juslin et al (2000) reinforce the position by examining the hard-easy effect in their paper. They find that binary choice questions give little to no support for a cognitive-processing bias. There is a near elimination of the hard-easy effect when there is control for scale-end effects and linear dependency (Juslin et al 2000). As a consequence of the aforementioned, Klayman and Soll (2003) attempt to tease out the unsystematic judgmental error contributing to confidence in their experiments. They find that with binary questions overconfidence nearly evaporates when testing a representative set of questions from multiple domains. With subjective intervals, however, they observe as much as 45% overconfidence even with representative sets of questions (Klayman and Soll 2003). The main findings are as follows: 1) unsystematic variation contributes to overconfidence, but the main cause is subjective intervals being too narrow; 2) the format by which subjective intervals are solicited has a large effect on overconfidence and interval size; and 3) men are more overconfident than women (Klayman and Soll 2003).

On the topic of subjective intervals, Kirchler and Maciejovsky (2002) set up a controlled experiment simulating asset markets to test subjects overconfidence in a finance setting. They hypothesize that, 1) subjective confidence intervals are too narrow, as they exclude observed trading prices too often, 2) traders are overconfident with respect to accuracy of their price predictions, 3) overconfidence starts low and increases with time (trading periods), and 4) trading volume is negatively correlated with individual earnings (Kirchler and Maciejovsky 2002). They find support for hypothesis one and three. However, Kirchler and Maciejovsky (2002) find subjects both over- and underconfident, and also well-calibrated in other cases, lending no support to the second hypothesis . They also find that trading volume is, on the contrary, positively correlated to individual earnings.

In the attempt of mitigating overconfidence, two papers emerge early on from Lichtenstein and Fischhoff (1980), and Arkes et al (1987). Lichtenstein and Fischhoff (1980) train subjects in item tasks followed by intensive performance feedback on calibration. They find that giving feedback to subjects, that were not well-calibrated to begin with, is sufficient enough to improve calibration. Arkes et al (1987) apply a different approach. They do not include any hands-on training, instead they provide subjects with feedback on accuracy or tell them that they will engage in a group discussion after the questionnaire. In both experiments Arkes et al (1987) find that the experimental subjects display lower levels of overconfidence compared to the control group. They conclude that less direct attempts to improve calibration, as research shows, may have an advantage over more direct approaches (Arkes et al 1987).

In contrast to the aforementioned findings, Pulford and Colman (1997) give subjects immediate feedback on general knowledge questions without success. Instead they conclude that feedback seems to be effective in improving calibration when questions are consistently difficult, not for medium-difficult or easy questions.

A more recent study from Speirs-Bridge et al (2010) seeks out to reduce overconfidence in interval judgments of experts. The researchers make use of the approach utilized by Klayman and Soll (2003). Klayman and Soll (2003) find that the question format in interval estimates influence the degree of overconfidence. In a 3-point format, subjects are asked to provide a lower and upper bound, and also a "best guess" serving as the median of the interval. Speirs-Bridge et al (2010) take it one step further and develop a 4-step process. As a fourth step, subjects are asked to assign a level of confidence themselves. Most other studies provide the level of confidence as a given or within a range. The authors conclude across three different experiments that the 4-step procedure results in minimal overconfidence of an average 11.9% (Speirs-Bridge et al 2010). They also find the 4-step process superior to the 3-step process for interval estimates for containing the correct answer.

Another recent study investigates debiasing methods in real-life settings for achieving lasting improvement (Aczel et al 2015). The study includes a list of ten biases, overconfidence being one of them. The researchers make use of Larrick's (2004) debiasing method that is divided into three categories, technological, motivational, and cognitive. Following debias training of subjects, the paper concludes that training only shows improvements for statistical biases such as Base rate neglect and Regression to the mean (Aczel et al 2015). No improvement is shown for biases such as Anchoring, Framing effect, and Overconfidence among others.

The current literature on overconfidence, and the ways to deal with its effects, come with both mixed results and ambiguity. It comes as no surprise that working with a phenomenon that constitutes human behavior gives rise to uncertainty. The intent of the current study is to broaden the spectra with respect to the uncertainty. The premise is that subjects do exhibit overconfidence even in binary format. Support is obtained for this claim through the experiment. The next mission is to link it up to interval estimates and the confidence expressed in too narrow intervals. This is a novel task, and especially, in addition to giving feedback on calibration preceding the interval estimates. Consequently, this opens up a new avenue for research to explore overconfidence and ways to mitigate its effects.

3. Overconfidence

3.1 Models to Measure Overconfidence

For the purpose of this paper, the definition of overconfidence will be situated from the viewpoint of Ackert and Deaves (2010) course literature on Behavioral Finance. *Overconfidence is the tendency for people to overestimate their knowledge, abilities, and the precision of their information, or to be overly sanguine of the future and their ability to control it* (Ackert and Deaves 2010, p.106). The different ways to measure overconfidence will be elaborated on separately in this section. The relevant measures utilized in this study for overconfidence include Interval Estimates, Miscalibration, and Better-than-average effect.

Interval Estimates

Interval estimation studies measure overconfidence in terms of confidence intervals. The main approach is quite straightforward. Usually, subjects are given a question where they provide the answer within a given range or interval. For example, one might ask *How many malls do you think exist in Sweden?*. The subject is then encouraged to provide an estimated lower- and upper bound with X % accuracy/confidence. It could look something like this, *I am 90 % certain that the answer lies somewhere between* ___X___ *(lower bound) and* ___X___ *(higher bound).* Both Klayman and Soll (2003), and Juslin et al (1999), postulate that this type of answer can be modeled as two binary decisions with 95 % certainty of the lower limit and 95% certainty of the upper limit. The main interest of interval estimates is to see to what extent the spread of intervals can predict the confidence level of subjects. In other words, is my assertion of the given range combined with X% certainty compatible with the narrowness or width of my interval? A way for researchers to answer this question is to take the estimated probability, say a confidence of 50%, and subtract the average percentage of intervals containing the correct answer to observe an effect. If the sample group provides an average hit rate of 40% then the group as a collective display overconfidence by 10% (50% – 40%). Simply put in equation format one can conclude the following:

Asserted Confidence Level – Average hit rate = X (underconfidence if $X < 0$ and overconfidence if $X > 0$)

This setup opens up for its own limitations questioning factors that may or may not contribute to levels of bias. Several aspects need to be taken into account for correctly assigning the findings to bias rather than noise in the data. This does, however, give a crude estimation of the level of overconfidence. This study applies a different approach. The median of the aggregated intervals is used as the cutoff for too narrow intervals. More on this in the Method section.

Miscalibration

The notion of calibration and interval estimates is at times used interchangeably throughout the literature. However, throughout this work the two will be kept separate and expanded upon on as such.

Miscalibration arises from the interaction between asserted accuracy and outcome. It is the tendency for people to over- or underestimate their precision of knowledge (Ackert and Deaves 2010). The conventional way of testing for calibration is to provide subjects with knowledge-based questions, and subsequently ask for the number of questions they believe to have answered correctly. If a subject overestimates their outcome they are deemed overconfident, and vice versa (Bhandari et al 2008). The same principle applies in this study. For the knowledge questions in the calibration part, subjects estimated score and their actual score will represent their calibration.

Better-than-average effect

The better-than-average effect is a self-explanatory term in and of itself. Studies have shown that when people are asked to rate themselves relative to the average person, they find that a disproportionate amount of people rate themselves as above average (Ackert and Deaves 2010). A classic experiment question that sheds light on the better-than-average effect is when participants are asked to rate their driving skills in relation to their peers. Svenson (1981) find that in his sample 88% from the US group, and 77% from the Swedish group, of participants rate themselves as safer drivers than the median. This type of pattern is shown time and time again among self-assessed attributes. This serves as the basis of what's referred to as the better-than-average effect. The better-than-average effect is used in this study as an extra dimension to investigate overconfidence in the sample as a whole.

3.2 Hypothesis

It is likely to think that individuals can learn from their past experience and mistakes. To overestimate ones knowledge is not ideal. Therefore, it is expected that if individuals get feedback on their imprecision that they will want to learn from it. To provide feedback could be a way to attenuate overconfidence, at least in the short run. Consequently, the hypothesis for this paper follows. Subjects that receive feedback on their overconfidence in the calibration part will exhibit less overconfidence in the interval estimates part.

4. Method

4.1 Research Method

- Quantitative
- Experimental design
- Survey-based

The method for conducting this study is in the form of an experimental approach. To investigate the proposed hypothesis, a survey-based experiment is carried out to collect the data necessary in order to extract any finding. The different elements that constitute the experiment will be described and discussed in this section.

4.2 The Experiment

The experiment aims at exposing subjects to a number of knowledge questions, and follow up with interval estimate questions. In this way the experiment is divided into two parts. This allows for elements of overconfidence to be present in both parts, first in terms of **calibration**, and secondly in **interval estimates**. At random, subjects receive feedback on the number of correct answers in the calibration part before moving on to the interval estimates part. The feedback is provided by a random generator regardless of how a subject scores, dividing subjects into a treatment group and a control group. This lends itself as a mechanism for aligning ones prior judgment of knowledge to be calibrated before the interval estimates part. This allows for the possibility to extract any effect from the hypothesis of feedback and self-correction.

4.3 Survey Design

The survey is generated using Qualtrics software. Subjects receive no information prior to taking the survey on it being an experiment of any sort. The survey is essentially divided into five subcategories: 1) introductory questions, 2) knowledge questions, 3) guesstimate, 4) interval estimates, 5) background information (see Appendix for excerpts). In the first part, subjects are asked whether they think it is important to bear general knowledge regarding pensions.

They are also asked to assert their level of knowledge regarding pensions. Three options are given to choose from, below-, above- or average knowledge. In the second part, subjects are provided with ten general knowledge questions/statements on the topic of pensions. The option to answer is given in binary form, true or false. In the third part, subjects are asked to estimate how many of the ten questions/statement they believe to have answered correctly. Following the guesstimate, a random generator will either display the score or proceed without. The random generator plays the role of dividing subjects into two groups, a control group and a treatment group. The ones receiving feedback become part of the treatment group, the others fall into the control group. To avoid selection bias, whether a subject is assigned to a control group or a treatment group is independent of how the subject has answered the questions from the first three parts. The fourth part includes two interval estimate questions. Question number one is related to exponential growth called "A penny doubling every day", with the exception of being in Swedish currency. One öre represents one cent. The question subjects are asked is "If you double a penny every day, how many days does it take to reach one million?". The second interval question is "How many gas stations exist in Sweden?". For the interval estimates, subjects are instructed to provide a range with a lower- and upper bound with 90% certainty that the true answer lie somewhere in between. The fifth, and final part of the survey, asks subjects for information regarding age, occupation, and level of education.

4.4 Control- and Treatment Group

	Treatment	Control
Overconfident	X	X
Not Overconfident		

The two-by-two above depicts what is of interest in this study. The boxes marked with an "X" are meant to illustrate where the focus lie. In the baseline regressions, I follow what has been described in the boxes by dividing subjects into a group who exhibit overconfidence and a group who do not exhibit overconfidence. In additional regressions, I further divide those subjects who do not exhibit overconfidence into two groups, a group who exhibit underconfidence and a group who exhibits neither underconfidence nor overconfidence (so-called calibrated).

The random generator, previously mentioned, serves as a bifurcation to divide subjects into a control group and a treatment group. The only difference between the two is the feedback given by the generator. Subjects in the control group also have to answer the same interval estimation questions. Thus, the only difference between the control group and the treatment group is that, before moving to part 4 to answer the interval estimation questions, only subjects in the treatment group get the feedback on how many correct answers they have received from the calibration part.

4.5 Data

The data is collected from distributing the survey via various email registries obtained from the University of Gothenburg. A final number of 171 full responses are collected. Some of the responses are not workable. For example, there are subjects that do not provide an interval but instead provide the same number twice. Furthermore, a threshold is set to adjust for extreme answers that ultimately will skew the data. For the interval question on how many days it will take for a penny doubling to reach a million, any response stating an upper bound of more than one million is taken out of the sample. The final sample size amounts to a total of 149 observations.

Variables

Table 4.1: Summary Statistics

Variable	Type	Min	Max	Mean	Median
Correct	Range	4	10	6.64	7
Guesstimate	Range	1	10	5.20	5
Confidence	Range	-7	3	-1.44	-1
Days	Interval	1	700,000	N/A	15
Gas Stations	Interval	1	10^{13}	N/A	2000

Correct is the number of correct answers in the calibration part.

Guesstimate is the number of correct answers the subject believed to have answered correctly in the calibration part.

Confidence is based on the subject's *Guesstimate* minus *Correct*. If a subject estimates 8 correct answers (*Guesstimate*) but in reality has 5 correct (*Correct*), then *Confidence* equals -3. If *Guesstimate* - *Correct* = 0, then the subject is considered *Calibrated*.

Days is the variable for the first interval question where subjects provide a lower- and a upper bound.

Gas Stations is the variable for the second interval question where subjects provide a lower- and a upper bound.

Table 4.2: Binary Variables

	"1"	"0"
Spread_1	77	72
Spread_2	76	73
Spread_X	39	110
Overconfidence	25	124
Underconfidence	103	46
Treatment	77	72

Spread_1 represents *Days* when the interval is too narrow. The cutoff for too narrow is taken from the median 15. When $Days \leq 15$, $Spread\_1 = 1$; otherwise $Spread\_1 = 0$.

Spread_2 represents *Gas Stations* when the interval is too narrow based on the median 2000. When $GasStations \leq 2000$, $Spread\_2 = 1$; otherwise, $Spread\_2 = 0$.

Spread_X is *Spread_1* multiplied by *Spread_2*. When $Spread\_1 = Spread\_2 = 1$, $Spread\_X = 1$; otherwise, $Spread\_X = 0$.

Overconfidence is based on *Confidence*. When $Confidence > 0$, $Overconfidence = 1$; otherwise $Overconfidence = 0$.

Underconfidence is also based on *Confidence* but in reverse. When $Confidence < 0$, $Underconfidence = 1$; otherwise $Underconfidence = 0$.

Treatment is the divide between treatment group and control group. If a subject receives feedback, $Treatment = 1$; otherwise $Treatment = 0$ (meaning that the subject is in the control group).

4.6 Logistic Regression

The aim is to investigate whether subjects that exhibit overconfidence in the calibration part, and receive feedback on it, will be less likely to provide too narrow intervals in the interval estimates part. Therefore, the variable of interest, which is also the dependent variable, is

Spread_X (see Regression Equation). Since Spread_X is dichotomous, the model to investigate the hypothesis falls on either a logit- or probit model. Logit- and probit models are similar in fashion with some key differences. One of these being that logit coefficients can be interpreted in odds ratios. Probit models can be applied to more advanced settings to account for heteroskedasticity among other things. This is not of necessity in the current set up for the experiment. Hence, a logit model is applied.

5. Results

5.1 Illustrative findings

Table 5.1 highlights the demographics in the sample group. As expected, the majority are educated and relatively young. This would be expected since the method of collecting data primarily came through email registries from university. As an indicative finding, it is helpful

Table 5.1: Sample Demographics

Male/Female	81/68
Born in 90s or earlier	83%
Students	78%
Employed	16%
University education	80%
Graduate level education	53%

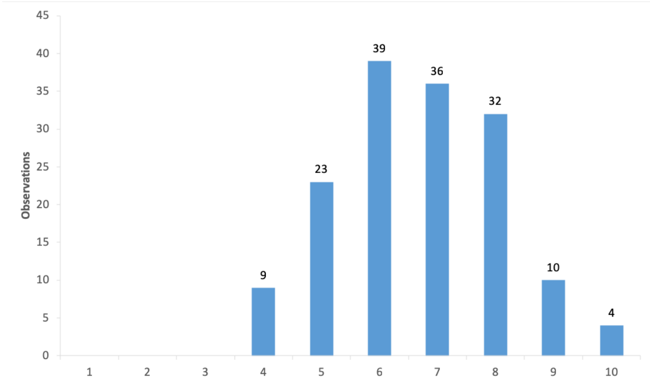
to begin by looking at Table 5.2 at a glance. It immediately becomes evident that the majority of subjects are underconfident rather than overconfident with respect to the calibration part. There are almost as many overconfident subjects as there are calibrated ones. It should be noted that there are relatively few subjects to work with in terms statistical inference.

Table 5.2: Categories of interest

	Overconfidence	Underconfidence	Calibrated	Total
Treatment	10	57	10	77
Control	15	46	11	72
Total	25	103	21	149

Figure 5.1 shows the distribution of *Correct*. An approximate 72% of respondents are found between scores of 6 to 8. The median of the sample group is 7. It suffices to say that the group as a collective show a relatively high score within this domain of knowledge. To build further on that note, almost half of the subjects, 48%, rate themselves as having average knowledge regarding pensions. About 18% of subjects rate themselves as above average. Such a distribution is in contrast to the better-than-average effect, leaving no support for it in this sample.

Figure 5.1: Distribution of *Correct*



Furthermore, Figure 5.2 displays the distribution of *Confidence*. Visually one can see that the sample is tilted towards underconfidence. The mean and median for the sample group is -1.44 and -1, respectively. It is clear that the sample as a collective exhibit underconfidence in the calibration part rather than overconfidence. It is also interesting to note the difference among subjects asserted level of knowledge, along with the aforementioned variables.

Figure 5.2: Distribution of *Confidence*

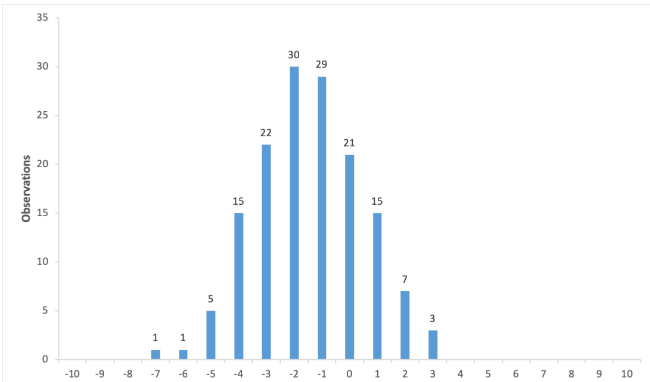


Figure 5.3 displays the disparity between knowledge groups and their performance. Albeit, the differences in *Correct* is in the decimals, it is true to say that the subjects displays a positive linear relationship across the three knowledge groups. It is interesting to note that differences between *Guesstimate* and *Correct* decrease as knowledge increase, although being underconfident. Ranging from below average to above average, the aggregated miscalibration in order of appearance is -2.24, -1.1, and -0.67. So, the more you claim to know the less of a miscalibration from this sample. Lichtenstein and Fischhoff (1977) find the same pattern but on the contrary, that all groups exhibit overconfidence.

Figure 5.3: Asserted level of knowledge (Mean)

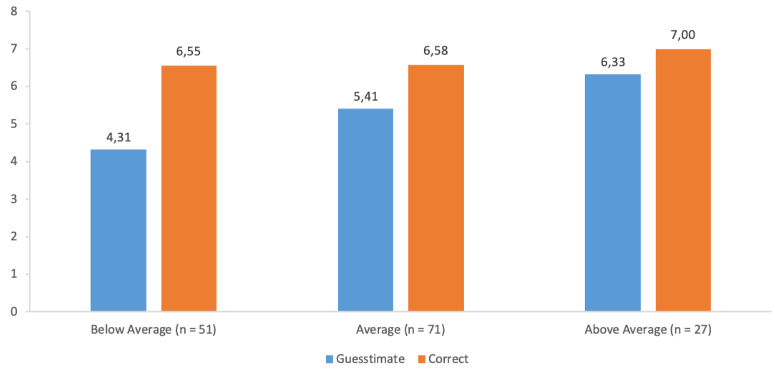
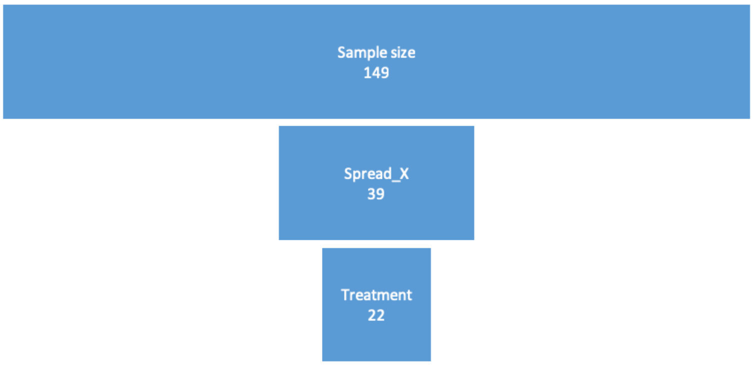


Figure 5.4 takes us to the interval estimates part of the experiment. Built on the premise of what constitutes narrow intervals, a total of 39 subjects provide too narrow intervals for the two questions combined. 22 out of these 39 are part of the treatment group. This gives an overview of the direction in the sample group that is analyzed. To keep in mind, this serves as a precursor to the logistic regressions that follow.

Figure 5.4: Subjects with too narrow intervals



5.2 Logistic Regressions

A total of six separate regressions are analyzed in order to investigate the working hypothesis. Three separate regressions are run in the format depicted by equation 5.1, and another three as depicted by equation 5.2. The dependent variables for equation 5.1 are Spread_X; Spread_1; Spread_2, representing too narrow intervals. The independent variables are *Overconfidence*, *Treatment* and *Overconfidence*Treatment*, respectively. β_0 serves as the intercept. It represents the benchmark in the regression equation. β_0 shows the probability of providing too narrow intervals for a subject that does not exhibit overconfidence, is not part of the treatment group, which also covers the interaction term for a subject that exhibits overconfidence in the treatment group. If all independent variables = 0, then β_0 is the probability for a benchmark subject to provide too narrow intervals, which is less than 1/2 as shown from the equation.

$$\Pr(*Y = 1) = \frac{\exp(\beta_0 + \beta_1\text{Overconfidence} + \beta_2\text{Treatment} + \beta_3\text{Overconfidence*Treatment})}{1 + \exp(\beta_0 + \beta_1\text{Overconfidence} + \beta_2\text{Treatment} + \beta_3\text{Overconfidence*Treatment})} \tag{5.1}$$

*Y = Spread_X; Spread_1; Spread_2

If the other independent variables = 0, the coefficient β_1 shows us the value for a subject that exhibits overconfidence for the probability of providing too narrow intervals. The sign of this coefficient is expected to be positive as an overconfident subject is expected to run a higher risk of providing too narrow intervals. If the other independent variables = 0, the coefficient β_2 shows us the value for a subject in the treatment group for the probability of providing too narrow intervals. The expectation for β_2 is to be close to zero. This encapsulates a subject that does not exhibit overconfidence, since the other variables = 0, and therefore is expected to have a marginal effect on the chances of providing too narrow intervals. The coefficient β_3 is of main interest as it captures the causal effect of an overconfident subject in the treatment group, given feedback, on providing too narrow intervals. If feedback can attenuate overconfidence then the sign of β_3 is expected to be negative.

$$\Pr(*Y = 1) = \frac{\exp(\beta_0 + \beta_1\text{Overc.} + \beta_2\text{Treat.} + \beta_3\text{O*Treat.} + \beta_4\text{Underc.} + \beta_5\text{U*Treat.})}{1 + \exp(\beta_0 + \beta_1\text{Overc.} + \beta_2\text{Treat.} + \beta_3\text{O*Treat.} + \beta_4\text{Underc.} + \beta_5\text{U*Treat.})} \tag{5.2}$$

*Y = Spread_X; Spread_1; Spread_2

Equation 5.2 serves as an extension of equation 5.1 with three additional regressions. Two independent variables are added, *Underconfidence* and *Underconfidence*Treatment*. β_0 shows the probability of a benchmark subject that does not exhibit overconfidence nor underconfidence, and is not in the treatment group, for providing too narrow intervals. This is essentially a subject that is calibrated (*Confidence* = 0). For β_1, β_2, and β_3, the expectations are the same as they are in equation 5.1. It is expected for β_1 to be positive, β_2 to be close to zero, and β_3 to be negative. If all other independent variables = 0, β_4 shows a subject that exhibits underconfidence for the probability of providing too narrow intervals. It is expected for the sign of this coefficient to be negative or close to zero, since it is not expected for underconfidence to show positive effects for the probability of providing too narrow intervals. The coefficient β_5 shows a subject that exhibits underconfidence in the treatment group for the probability of providing too narrow intervals. The expectation is for β_5 to be either negative or near zero, since an underconfident subject in the treatment group is not expected to show a positive effect for the probability to provide too narrow intervals.

Table 5.3 shows results from the logistic regressions. The restricted version of the depen-

Table 5.3: Dependent Variable(s): Too narrow intervals

Variable	Spread_X	Spread_X	Spread_1	Spread_1	Spread_2	Spread_2
Overconfidence	1.025	1.897	0.658	0.875	0.042	0.426
	(0.62)	(1.17)	(0.61)	(0.82)	(0.58)	(0.81)
Treatment	0.504	0.105	-0.125	-0.665	0.446	0.154
	(0.43)	(1.49)	(0.36)	(0.92)	(0.36)	(0.90)
O*Treatment	-0.946	-0.547	-0.163	0.377	-0.313	-0.021
	(0.97)	(1.72)	(0.92)	(1.25)	(0.89)	(1.22)
Underconfidence		1.021		0.269		0.473
		(1.11)		(0.67)		(0.69)
U*Treatment		0.402		0.613		0.324
		(1.56)		(1.00)		(0.98)
Constant	-1.431***	-2.303**	0.035	-0.182	-0.176	-0.560
	(0.34)	(1.05)	(0.26)	(0.61)	(0.27)	(0.63)
N	149	149	149	149	149	149

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

dent variable Spread_X, with respect to the signs of the coefficients, is largely in line with the hypothesis. As depicted by the constant, a benchmark subject that does not exhibit overconfidence and is not in the treatment group shows a negative sign on the probability of providing too narrow intervals. Compared to the benchmark, a subject that exhibits overconfidence and is not part of the treatment group shows an added 1.025. For the interaction term, a subject that exhibits overconfidence and is in the treatment group shows an added -0.946 + 0.504 + 1.025 = 0.583, compared to the benchmark. This is in the desired direction for the hypothesis and shows less added than for a subject that exhibits overconfidence and is not part of the treatment group for the probability of providing too narrow intervals. Looking at the extended version of Spread_X, we see the same pattern. A subject that exhibits overconfidence and is not part of the treatment group shows an added 1.897, compared to the benchmark subject that shows -2.303 for the probability of providing too narrow intervals. A subject that exhibits overconfidence and is part of the treatment group shows -0.547 + 0.105 + 1.897 = 1.455 for the probability of providing too narrow intervals, compared to the benchmark. This is again in the desired direction, showing less for a subject that exhibits overconfidence and is not in the treatment group. A subject that exhibits underconfidence and is not in the treatment group also adds to the probability of providing too narrow intervals by 1.021, compared to the benchmark. If a subject exhibits underconfidence and is part of the treatment group then 0.402 + 1.021 + 0.105 = 1.528 is added for the probability of providing too narrow intervals, compared to the benchmark. Despite 1.528 being less than for a subject that only exhibits overconfidence with 1.897, this is not in the desired direction.

Looking at the regression when Y = Spread_1 we can see some different results. In the restricted version, a subject that exhibits overconfidence and is not part of the treatment group shows an added 0.658 to the probability of providing too narrow intervals, compared to the benchmark of 0.035. A subject that exhibits overconfidence and is part of the treatment group shows -0.163 -0.125 + 0.658 = 0.37 added to the probability of providing too narrow intervals, compared to the benchmark. This is in the desired direction that the probability is less for a subject that exhibits overconfidence and is not part of the treatment group that shows 0.658. The extended version of Spread_1 tells a different story. A subject that exhibits overconfidence and is part of the treatment group shows 0.377 - 0.665 + 0.875 = 0.587 for the probability of providing too narrow intervals, compared to the benchmark. The interaction term for *Overconfidence*Treatment* is now positive, which is not in the desired direction. Although, the value is smaller for a subject that exhibits overconfidence and is in the treatment group than for a subject that exhibits overconfidence and is not in the treatment group, with 0.587 and 0.875, respectively.

Looking at the restricted regression when Y = Spread_2 more results are observed. A subject that exhibits overconfidence and is not part of the treatment group shows an added 0.042 for the probability of providing too narrow intervals, compared to the benchmark of -0.176. A subject that exhibits overconfidence and is part of the treatment group shows -0.313 + 0.446 + 0.042 = 0.175 for the probability to provide too narrow intervals, compared to the benchmark. This is a higher value than for a subject that simply exhibits overconfidence and is not part of the treatment group, with 0.175 - 0.042 = 0.133. This is not in the desired direction of supporting the hypothesis. Similar results are observed when *Underconfidence* is controlled for in the the extended version. Here a subject that exhibits overconfidence and is not part of the treatment group shows an added 0.426 for the probability to provide too narrow intervals, compared to the benchmark -0.560. A subject that exhibits overconfidence and is part of the treatment group shows an added-0.021 + 0.154 + 0.426 = 0.559 for the probability to provide too narrow intervals. This is an increase in the value compared to a subject that exhibits overconfidence and is not part of the treatment group with 0.559 - 0.426 = 0.133. Also, the signs of the coefficients for *Underconfidence* and *Underconfidence*Treatment* are both positive. A subject that exhibits underconfidence and is part of the treatment group shows an added 0.324 + 0.473 + 0.154 = 0.951 for the probability to provide too narrow intervals. This is a larger than for both previously mentioned values, for a subject that exhibits overconfidence in the treatment group and not in the treatment group, with 0.559 and 0.426, respectively.

A caveat for the regression output is that there are only two variables that show significance from the entire layout. These being the constant in both the restricted and extended version when Y = Spread_X. This renders statistical inference without power. For some regressions the desired effects are observed given the interaction with feedback. However, results go in both directions for other regression, which undermine the efficacy of feedback. For example, when Y = Spread_2 and *Underconfidence* is controlled for, it is problematic that the value for a subject that exhibits underconfidence in the treatment group is larger than one that exhibits overconfidence in the treatment group. Nonetheless, the results become difficult to draw an inference from in the absence of statistical power, which causes for weak evidence in support of the hypothesis.

6. Discussion

In terms of statistical power, the experiment does not provide sufficient significance to enable robust inference. This leads to inconclusive support for the working hypothesis. The majority of subjects exhibit underconfidence in terms of miscalibration. This could be a contributing factor to the incomplete analysis for attempting to predict the efficacy of feedback to interval estimates narrowness.

One third of subjects provide too narrow intervals for the two interval estimates combined. In light of crude measures, such as the use of median to divide between too narrow or not, one can ask what constitutes too narrow of an interval? However, it should be noted that the aim is to investigate the effects of feedback on narrowness between subjects based on prior overconfidence in calibration. This as well opens up for criticism. Could it be the case that a subject's calibration is not fully representative of its confidence? It may very well be so. There is no shortness of debate regarding this matter. Some argue that perfect precision is an unattainable ideal. On the other hand, others argue that the observation of imperfection can be worked on and improved. I belong to the second camp. Although, one cannot avoid questioning the ambiguity of calibration as confidence. Or if the emergence of miscalibration is significantly different from noise in the data.

Nevertheless, it is worth noting that 10% of interval estimates combined for both questions contained the right answers. This raises the issue with too narrow intervals. 26% of interval estimates contained the true answer for Spread_1, and 44% for Spread_2, respectively. This is in addition to asking subjects to provide intervals with 90% certainty. Like other studies have shown, subjects do as a collective provide too narrow intervals. At least in terms of the degree of certainty relative to how many intervals contain the true answer. Several explanation have been suggested for why this is the case. Some researchers allude to the weighting function held up in Prospect Theory. Subjects overweight low probabilities, and simultaneously, underweight high probabilities (Kahneman and Tversky 1979). The framing of the questions can also play a part. Like Klayman et al (1999) show, how you ask the question can significantly impact the outcome.

Despite the lack of significance in the regressions, there are findings to be highlighted. The signs of the coefficients point in the expected directions for some findings coupled with feedback, mainly when Y = Spread_X. It may be that feedback does, in fact, have a conclusive impact on attenuating overconfidence. However, such an inference is not fully supported based on the findings of this paper.

Limitations

There are various limitations that ought to be taken into consideration when evaluating the data. First and foremost, the experimental setting was under no circumstances a controlled setting. This can impair the data significantly. Subjects can search for answers while filling out the survey. There are no constraints on resuming the survey if chosen to exit in the middle of it. Randomness is another aspect to take into account. Anyone who attempts to draw conclusions from binary answers faces the challenge of randomness. Granted, subjects are asked to note how many answers they believe to have answered correctly, which would be calibrated for. However, it becomes a challenge to eliminate the variable of randomness when the calibration is the variable of interest. Furthermore, the risk of sampling bias is present at all times. Having mainly students respond to the survey can lead to misrepresentation. The sample size itself can also come under scrutiny for the same reason.

7. Conclusion

Summary

This paper examined whether feedback could attenuate overconfidence. To answer this question an experiment was carried out to test the hypothesis in a survey-based fashion. Although signs indicating that feedback can have an impact on a subject's overconfidence, the evidence in support of this is too weak to draw a conclusion. In accordance to earlier findings, it remains that subjects interval estimates are not in line with provided degree of certainty. On providing intervals with 90% certainty, only 10% of subjects interval estimates contained the true answer for both questions. This further supports the claim that subjects provide too narrow intervals with given confidence levels. In contrast to overconfidence on the part of interval estimates, the majority of subjects exhibit underconfidence in terms of calibration. This further implicates the notion of overconfidence as an impartial concept restricted to one domain. Further research could provide a clearer understanding of the inner workings behind the psychology of decision-making in this area.

Future Research

It would be interesting for future research to elaborate on feedback as a potential mechanism to attenuate overconfidence. A more thorough experiment design can approach the issue from a different angle. Iterating feedback on performance over different domains of knowledge might show a lasting effect in the decision-making process. Furthermore, new measures of overconfidence could be defined and cross-examined to corroborate whether the phenomenon is consistent or unsystematic and domain-specific. For individuals to show overconfidence in one area and simultaneously underconfidence in another undermines the notion. Perhaps future research can come to terms with where and how overconfidence is present. And if necessary, rule it out as one coherent idea if the pattern of behavior lacks clarity.

8. Bibliography

[1] Ackert, L, F. Deaves, R. (2010). Behavioral Finance: Psychology, Decision-Making, and Markets. Mason, Ohio: South-Western Cengage Learning.

[2] Aczel , B., Bago, B., Szollosi, A., Foldes, A., Lukacs, B. (2015). Is it time for studying real-life debiasing? Evaluation of the effectiveness of an analogical intervention technique. *Frontiers in Psychology, 6*, 1120. doi: 10.3389/fpsyg.2015.01120

[3] Arkes, H. R. (1991). Costs and Benefits of Judgment Errors: Implications for Debiasing. *Psychological Bulletin, 110(3)*, 486-498. doi: 10.1037/0033-2909.110.3.486

[4] Arkes, H. R. Christensen, H. Lai, C. Blumer, C. (1987). Two Methods of Reducing Overconfidence. *Organizational Behavior and Human Decision Processes, 39*, 133-144.

[5] Bhandari, G. Hassanein, K. Deaves, R. (2008). Debiasing investors with decision support systems: An experimental investigation. *Decision Support Systems, 46*, 399-410. doi: 10.1016/j.dss.2008.07.010

[6] Barber, B. Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *Journal of Finance, 55*, 773–806.

[7] Buehler, R. Griffin, D. Ross, M. (1994). Exploring the "Planning Fallacy": Why People Underestimate Their Task Completion Times. *Journal of Personality and Social Psychology, 67(3)*, 366-381. doi: 10.1037/0022-3514.67.3.366

[8] De Bondt, W. Thaler, R. (1987). Further Evidence on Investor Overreaction and Stock Market Seasonality. *The Journal of Finance, 42(3)*, 557-881.

[9] Erev, I. Wallsten, T.S. Budescu, D.V. (1994). Simultaneous Over- and Underconfidence: The Role of Error in Judgment Processes. *Psychological Review 101(3)*, 519–528. doi: 10.1037/0033-295X.101.3.519

[10] Gigerenzer, G. (1991). How to Make Cognitive Illusions Disappear: Beyond "Heuristics and biases". *European Review of Social Psychology , 2*, 83–115.

[11] Griffin, D. Tversky, A. (1992). The weighing of Evidence and the Determinants of Confidence. *Cognitive Psychology, 24*, 411-435. doi: 10.1016/0010-0285(92)90013-R

[12] Juslin, P. Wennerholm, P. Olsson, H. (1999). Format Dependence in Subjective Probability Calibration. *Journal of Experimental Psychology: Learning, Memory and Cognition, 25(4)*, 1038-1052. doi: 10.1037/0278-7393.25.4.1038

[13] Juslin, P. Winman, A. Olsson, H. (2000). Naïve Empiricism and Dogmatism in Confidence Research: A Critical Examination of the Hard-Easy Effect. *Psychological Review, 107(2)*, 384-396. doi:10.1037//0033-295X.107.2.384

[14] Kahneman, D. Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *The Econometric Society, 47(3)*, 263-291.

[15] Kirchler, E. Maciejovsky, B. (2002). Simultaneous Over- and Underconfidence: Evidence from Experimental Asset Markets. *The Journal of Risk and Uncertainty, 25(1)*, 65–85.

[16] Klayman, J. Soll, J, B. González-Vallejo, C. Barlas, S. (1999). Overconfidence: It Depends on How, What, and Whom You Ask. *Organizational Behavior and Human Decision Processes, 79(3)*, 216-247. doi: 10.1006/obhd.1999.2847

[17] Klayman, J. Soll, J, B. (2003). Overconfidence in Interval estimates. *Journal of Experimental Psychology, 30(2)*, 299-314.

Larrick, R. P. (2004). Debiasing. In Koehler, D. and Harvey, N. Blackwell Handbook of Judgment and Decision Making. Blackwell Publishing.

[18] Larrick, R. P. (2004). Debiasing. In Koehler, D. and Harvey, N. Blackwell Handbook of Judgment and Decision Making. *Blackwell Publishing.*

[19] Lichtenstein, S. Fischhoff, B. (1977). Do those who know more also know more about how much they know?. *Organizational Behavior and Human Performance, 20*, 159-183.

[20] Pulford, B. D. Colman, A. M. (1997) Overconfidence: Feedback and Item Difficulty Effects. *Personality and Individual Differences, 23(1)*, 125-133. doi: 10.1016/S0191-8869(97)00028-7

[21] Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers?. *Acta Psychologica, 47*, 143-148.

[22] Speirs-Bridge, A. Fidler, F. McBride, M. Flander, L. Cumming, G. Burgman, M. (2010). Reducing Overconfidence in the Interval Judgments of Experts. *Risk Analysis, 30(3)*, 512-523. doi: 10.1111/j.1539-6924.2009.01337

[23] Thaler, R. H. Sunstein, C, B. (2008). Nudge: Improving Decisions about Health, Wealth and Happiness. *United States of America: Yale University Press.*

[24] Thaler, R. H.Sunstein, C, B. Balz, J,P. (2010). Choice Architecture.

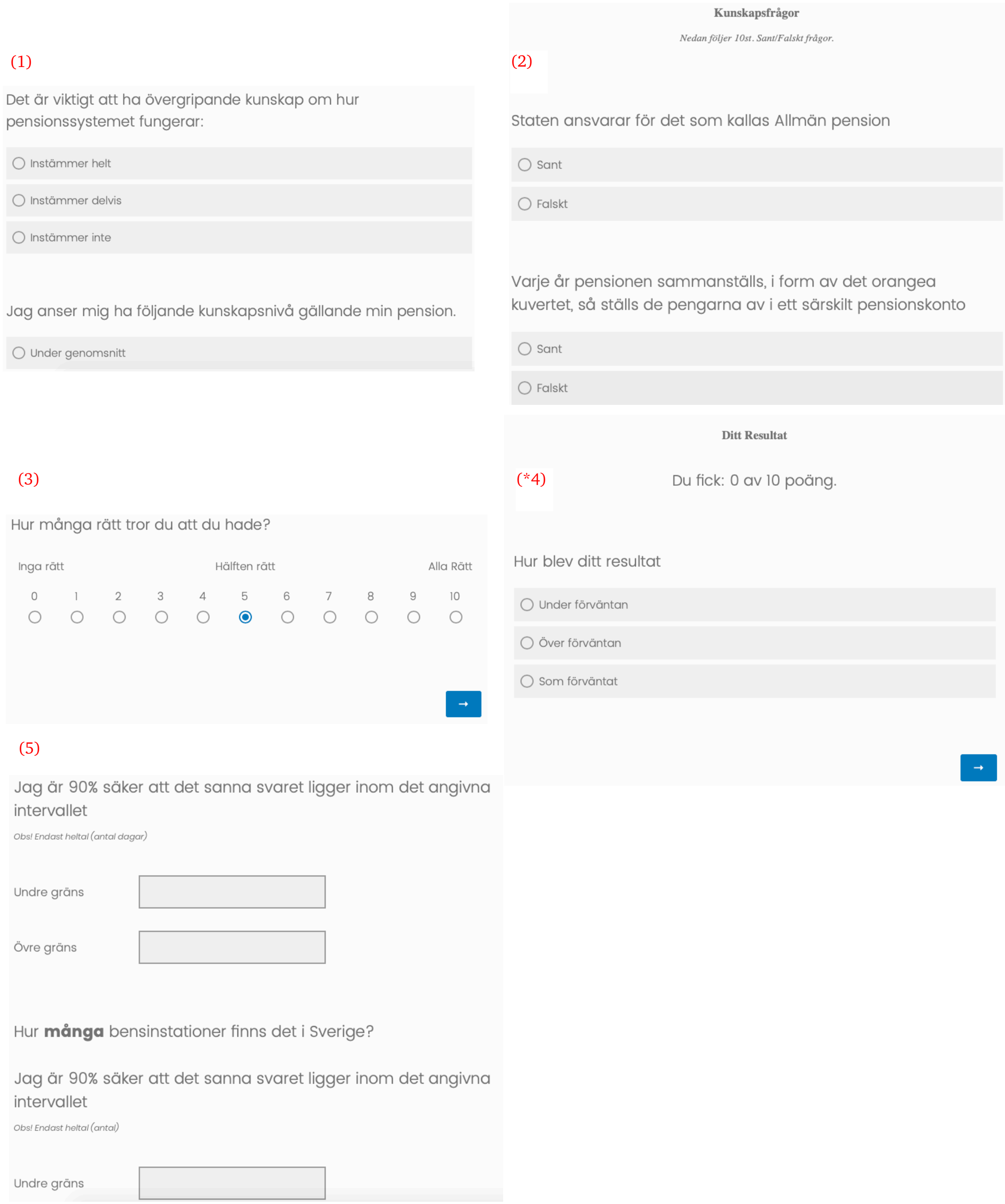
[25] Tversky, A. Kahneman, D. (1974). Judgement under Uncertainty: Heuristics and Biases. Science, 185, 4157, 1124-1131.

[26] Tversky, A. Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics, 106(4)*, 1039-1061. doi: 10.2307/2937956

[27] Winman, A. Juslin, P. Hansson, P. (2004). Subjective Probability Intervals: How to Reduce Overconfidence by Interval Evaluation. Journal of Experimental Psychology: Learning, Memory and Cognition, 30, 6, 1167-1175. doi: 10.1037/0278-7393.30.6.1167

8. Appendix

1) Survey Layout



*Only displayed to treatment subjects