Master Thesis Behavioural Economics

Can people be nudged into becoming better forecasters through a market environment?

Name: Eugenio Magni

Student number: 619342

Supervisor: prof. dr. Aurélien Baillon

Date of final version: 2/11/2022

Abstract

Given the historical academic interest in the power of markets as information aggregators and the importance of producing effective and reliable forecasts on everyday economic problems, it is important to assess the viability of prediction markets as forecasting tools. Specifically, we tested the effectiveness of a simulated prediction market against a simple opinion poll, to answer the research question: will the accuracy of the predictions expressed by the treatment group (market) be higher than those expressed by the control group (poll)? The results partially corroborate our premises, enlightening the possibility of using markets to improve the forecasting skills of a group of people, even though the results do not hold for all the formulated hypotheses.

Keywords: prediction markets, economic forecasting, superforecasting.

Acknowledgements

I am extremely grateful to prof. dr. Aurélien Baillon for his twofold role: first and foremost, he was the best supervisor one could hope for, providing impactful feedback as fast as humanly possible and being always available and kind; secondly, his hard work as Master Coordinator ensured a great academic experience even in these complex times of pandemic.

Many thanks to dr. Jan Stoop, who skilfully transmitted his passion for economic experiments in the course Experimental Economics and inspired this work.

Words cannot express my gratitude to my Iva, whose love and unconditional support pushed me through this endeavour, as well as her proof-reader skills.

I'm extremely grateful to my family, Sandro, Patrizia and Elisa, who supported me physically and mentally through my entire academic career.

Lastly, I would like to extend my sincere thanks to the Šimović family, who ensured that I felt welcomed in a foreign country and accepted me amongst their own.

Table of Contents

1 Introduction

Forecasting problems are permeating every field of economics and finance (and nearly every other) where the necessary decisions of agents depend on uncertain future values. Forecasting can be broadly defined as a process involving providing information on future values of one or more variables of interest (Elliott et al., 2008). If we consider a household, the simple decision of how much to save for the proverbial rainy day presumes the ability to forecast a stream of future incomes or losses. Considering firms, the decisions affecting the capital structure, such as when to invest and how to finance investments, depend on the managers' ability to forecast future cash flows (Elliott et al., 2008). Similarly, in the public sector, the decision affecting the level of pensions of a country depends on the ability to forecast future tax income and GDP growth. Academic interest in prediction and forecasting has grown steadily in recent decades (Tsiralis et al., 2007). Researchers have been trying to understand how to improve expert's prediction and if the use of market mechanisms can be implemented to enhance prediction capacities. This work sets out to investigate if a random pool of people can be nudged through a market-like situation into becoming better forecasters.

Economic forecasting provides forecasters and observers with quick and more useful feedback. Future outcomes of predicted variables are observed within a reasonable amount of time, revealing whether the forecasting performance was poor or not. This work tries to test the stream of literature that looks to improve predictions (even arriving to the creation of teams of superforecasters) using market frameworks, exploiting the informational power of markets which is one of the most recognised economic forces of history. Since the works of Hayek and the rest of the Austrian School, economists have promoted the power of markets as an excellent mechanism to collect and express aggregated information normally dispersed between several actors. The classic efficient market hypothesis entails that the latest price on a market reflects all information available to market participants. It is possible to incorporate this characteristic in a prediction market: the current latest price in a prediction market is thought to reflect the participants' current best guess on the probability of the event (Atanasov et al., 2016).

This work exploits the properties of economic forecasting to set up an experiment with a control and a treatment group. The main research question of this thesis regards the increase in accuracy derived from the use of prediction markets: is the accuracy of the predictions expressed by the treatment group higher than those expressed by the control group? A survey asking about the likelihood of ten future uncertain events (but resolving within a month from the start of the survey) is handed out to both groups. The questions are the same but, while the control group has the questions framed like a simple poll, the treatment group received the questions framed as a small-scale simulation of a prediction market.

The next section covers the relevant past literature, detailing how this thesis can fit in the stream of literature on prediction markets and the improvement of forecasting applied to an economic setting.

2 Literature review

2.1 (Super) Forecasting

As this work bases its core on the purpose of improving forecasting for economic applications, the literature on economic forecasting and, consequently, belief elicitation is essential to create the methods for the experiment. Forecasting problems are present in all areas of economics (and everyday life in general), especially where people have to act based on the uncertain future value of one or more variables of interest. For example, a family has to decide how much to save up every month of the family budget (prediction of the stream of future wages and return on savings) and a company has to choose how much and when to invest (prediction of future cash flows and capital structure) (Elliott et al., 2008). Another major field of application is the public sector, which often requires accurate predictions within a timeframe of decades or centuries (think of the construction of major public infrastructure, such as a bridge or a highway). A unique feature of this kind of forecasting is that often its accuracy is revealed within a reasonable time period and that a prediction can be dynamically changed as more data on the matter become available (Elliott et al., 2008). Once a forecast has been made, it is important to understand how accurate it was, so that it can be reliably incorporated as a model into everyday life decisions. To understand the importance of forecasting in the economic context, consider two examples: on the microeconomic level, an investor's portfolio allocation decisions; on the macroeconomic level, Central Banks' predictions of future inflation and their consequent interest rates decisions. Forecasting can be broadly defined as the process involved in providing information on future values of one or more variables of interest. It is worth mentioning that in the case of the possibility of forecasting at scale (Taylor et al., 2017), with a high volume of data available, automated tools (algorithms) for forecasting are viable. However, it is difficult to tune them precisely enough to incorporate the greatest tools of human cognition, which are intuition and heuristics.

An interesting contribution to the same problem comes from Hogarth et al. (1981): they assert the importance of formally structuring and incorporating forecasting and planning (F&P) into the everyday workflow of businesses, non-profit and public organizations. F&P involves a formal framework to channel human intuition and judgemental abilities. As human ability to process information is limited and people adapt easily to dynamic changes, a mechanism capable of efficiently summarizing complex information is desperately needed.

2.2 The Good Judgement Project

This research is primarily driven by the seminal work of the Good Judgement Project, which laid the foundation for the development of the concept of superforecasters (Mellers et al., 2015). Over the span of two years, they selected through a tournament of predictions on real-life geopolitical events the top performers and assigned them to small teams, which also underwent a focused debiasing training. The results show significant outperforming of "regular" experts or randomly assigned teams, demonstrating the power of exceptional talent selection, debiasing training and intrinsic motivation of experts. Superforecasters seem to possess specific cognitive abilities that can be enhanced with well-aimed training (Fong et al., 1986). Better preparation, an enriching environment, intrinsic motivation (mostly enjoyment derived from solving problems) and open-mindedness (superforecasters are more open to being challenged on their opinions and beliefs) are the pillars of the success of these superforecasters (Schoemaker et al., 2016). This study took a substantial amount of academic manpower and a long time, to find a small number of superforecasters from a big pool of subjects.

Another paper elaborating on the Good Judgement Project is the work from Atanasov et al. (2016), which has the revealing title of "Distilling the wisdom of the crowds". It is the main inspiration for the experiment performed in this work, as it created a large-scale experiment comparing the prediction accuracy of a prediction market and a prediction poll. In particular, the experimenters could count on 2400 subjects participating in two years of geopolitical prediction tournaments and their performances were assessed with the Brier score. Their results corroborate the fact that, with a sufficient number of participants, prediction markets have an edge on simple polling. However, it is notable that opinion polls have an important property that is precluded to prediction markets, which is the possibility of keeping predictions private. The market price reveals the average aggregated prediction to all participants and, in some cases, this could be detrimental (for example in the context of commercial predictions for the inside use of the management of a company). It is possible to collect predictions through an opinion poll without revealing aggregate predictions to the other participants.

An important contribution is given by Katsagounos et al. (2021): they tried to replicate this study with a much smaller starting pool of subjects, arguing that most companies do not have the resources to adopt such a technique to improve the forecasting results of their employee. The results of their study, limited in scale and time, corroborate the GJP line of reasoning, while adding that a reward scheme may be useful for keeping engagement high.

2.3 Prediction markets

A second line of seminal research for this thesis is the literature on prediction markets, a topic that started to be explored at the end of the XX century and saw a sharp rise in production of academic papers in the last decades (Tziralis et al., 2017). Prediction markets, also known as information markets, can be defined as markets in which participants trade securities that offer returns on contingent on the occurrence of an event. As more traditional markets, they assume that prices are information containers (Hayek, 1945), in this case efficiently reflecting and aggregating all relevant information about the outcomes of a specific event. This can only happen if, collectively, market participants possess sufficient knowledge to predict an event with certainty.

However, there is a clear difference between stock and information markets, as the latter have a unique characteristic: accuracy. It is defined as the market's ability to predict the resolution of an event correctly, which is revealed once the event resolves. This is not present in traditional stock markets, as they only try to assess the true value of a firm, for a continuous and infinite amount of time. As stock markets, prediction markets have limitations. The literature (Wolfers et al., 2004) pinpoints three main problems: market failures, arbitrage and ethical degrading.

The most classical market failure is the emergency of incentives to "bet against the market" or, in other words, speculation. As opposed to traditional markets, prediction markets do not place constraints on short-selling and operate on a small scale, eliminating the main causes of speculative bubbles. A second problem that markets face is arbitrage. Once again, prediction markets seem to offer significantly more obstacles to arbitrageurs, because of the small scale and lesser predictability of price patterns. The last problem is more difficult to tackle, even for prediction markets: Falk et al. (2013) show how implementing market conditions in decision-making heavily erodes moral standards of participants. Although this is surely a problem, the strongest, profit-driven incentive that prediction markets provide is towards a correct prediction. This does not depend on the nature of the predicted event nor on the forecaster's opinion or moral evaluation of the matter, substantially limiting this problem regarding prediction markets. Although they have their limitations, prediction markets significantly outperform other methods of prediction such as experts' committees and opinion polls and are best fitted to be used as decision support systems (Berg et al., 2003).

An important point on the comparison between prediction markets and opinion polls as a decision support mechanism for organizations is brought up by Maciejovsky et al. (2020). As organisations often use teams or committees for decision making, it is important to organise an effective system for information-sharing and collective learning. These teams usually operate in face-to-face informal settings, which are markedly different from the highly computerised setting of prediction markets. Usually, the challenges presented to teams or prediction markets require the collective knowledge of all participants, meaning that participants should consider both shared and private information. It is established that markets are superior at pooling the information from all participants, because groups members tend to focus on shared information. This stems from the fact that market participants are incentivised to "put their money where their mouth is" (Hankins et al., 2011) and do not require perfect alignment between the organization and the employees' goals. Therefore, markets can be considered more efficient at information pooling and processing, but it has to be noted that forecasters tend to place more trust in groups than markets, mainly because prediction markets are not very well-known and widespread. Markets are also typically anonymous, allowing for people at lower ranks or positions in the organization to express their opinion more freely. Markets can accommodate a larger number of forecasters for longer time periods and are more difficult to be influenced, providing more valuable signals to management and decision makers than teams.

2.4 Belief elicitation and forecast verification

When making any kind of forecasting experiment, it is important to understand how beliefs of the participants can be elicited (Trautmann et al., 2011). The most important characteristic of an elicitation method (also known as *truth serums*) is that it must be incentive compatible. A mechanism is defined as incentive compatible if every participant can achieve the best outcome to themselves just by acting according to their true preferences. When economists prepare economical experiments, they usually assume that agents create beliefs rationally. In reality, agents may deviate from rationality. Various methods have been tested throughout the years to account for this phenomenon and for different risk stances, not only for the classic risk-neutrality case, but also to contemplate risk-aversion and risk-love. Trautmann et al. claim that more complex belief elicitation methods provide greater benefits than using simple direct asking, even though they come at a cost.

Forecast verification is the essential process of the *post ante* assessment of the accuracy of the prediction; it determines its quality and represents an essential component of any scientific forecasting system. The main bulk of research produced on the topic comes, originally, from weather forecasting: it is, in fact, the first field that was confronted by the problem of assessing the accuracy of a forecast with a binary (rain/no rain) or probabilistic outcome (with what probability is going to rain tomorrow?). Therefore, it is not a surprise that from this field come both the framework of forecast verification (Murphy et al., 1987) and a complete methodology to measure the accuracy of probabilistic predictions, the so-called Brier score, from the name of its developer (Brier, 1950). Forecast verification serves many purposes, including assessing recent trends in forecast quality, improving forecasting procedures and the forecasts themselves, and providing users with information needed to gain maximum benefit from the forecast. Verification methods vary based on the nature of the predicting variable, for example if it is categorical or probabilistic.

The forecasting verification process is required to meet the needs of a number of different groups, like forecasters, forecasting model developers and final users of forecast information. Possible uses include directing future research, effectively employing funding, assisting model developers in upgrading predicting models and, finally, supporting final users (companies and people) in the decision-making process. These are considered as 'forecasting stakeholders' (Casati et al., 2008). The most articulate forecasting method is indeed useless if it is not accompanied by a proper method that

measures, *ex post*, the accuracy of the predictions. Thus, it is important to cater the needs of every stakeholder when choosing which verification method to employ, to maximize the operational verification potential. The value of a forecast depends not only on how well it foreshadows future uncertain events but also upon its ability to be communicated to decision makers as well as their consequent capabilities to act accordingly. For most users, the quality of the verification measures is more important than the quality of the forecast itself. As a sidenote, it is useful to remind that forecasting and verification efforts always need to be proportional to the final user's financial sensitivity to the forecasted event, meaning costs and losses to be faced if a specific action is needed for a specific resolution of the uncertain event.

2.5 The Brier Score

One of the first complex methods for belief elicitation was developed by Glenn W. Brier (1950), from whom it took the name. Brier was a researcher working for the U.S. Weather Bureau in Washington, D. C. It was part of the literature strain that was trying to refine the technique of weather forecasting. At that time, it was of crucial importance to create a reliable scoring system, to evaluate the accuracy of weather predictions, especially about extreme events that can cause massive economic damage, like tornados and hurricanes. A verification system for forecasting needs to be first and foremost neutral, meaning that it should not influence the forecaster in any undesirable way. In particular, Brier devised a method to rate predictions with a binary, probabilistic outcome, meaning that it measures the accuracy of a prediction expressed with a probability (confidence statement) about an event with only two outcomes (ex. rain/no rain). The resulting score is calculated by taking the mean squared differences between the predictions and their corresponding event scores (more details and the precise formulas will be included in the Methods section of this paper). Thus, the lower the score, the greater the accuracy. The forecaster is incentivised to minimize his or her score by getting the forecasts exactly right or at least to state an unbiased estimate of the probability he or she cannot forecast perfectly. This verification system is not rewarding the use of extreme values (0 or 1) when they are not backed by extreme overconfidence in the prediction. In case of complete lack of prediction competences, the forecaster is still encouraged to state a probability (to the best of his abilities) rather than state the most frequent class on every occasion.

The Brier score is a good method for eliciting beliefs when agents are risk-neutral. However, this is often not the case, as previously discussed. The Binarized Scoring Rule (BSR), developed by Hossain and Okui (2011), is of particular interest. It is incentive compatible with irrespective of the risk-preference of the agent in question and it is valuable even if the agent is not an expected utility maximiser. It is similar to a quadratic scoring rule, but yields better performances. It creates a binary function that determines whether the agent will receive a fixed reward based on his or her own performance using an independently drawn random number. Experimental results of the BSR are really convincing, showing better results than any other method and using less assumptions. However, in the case of this thesis, the experiment does not present this specific problem, as the risk attitude of the participants do not affect the results, because they are not aware of the verification system before taking the survey, which is incentive-compatible because it is in the best interests of every participant, regardless of the risk attitudes, to try to express the best forecast possible and win the monetary prize for the best forecaster (see more in the Method section). Therefore, the Brier score is still the best choice for this experiment, as it is less complex and more reliable.

3 Research Hypotheses

This work tests, through an experimental design, if a market-like framing can improve the accuracy of forecasting compared to the control treatment - a simple survey. The research question that derives is: Can people be nudged into becoming better forecasters through a market environment?

Specifically, this study proposes to test one hypothesis, which is articulated in several sub hypotheses:

3.1 H1: The accuracy of the predictions expressed by the treatment group is higher than those expressed by the control group.

- H1.a: The average Brier score of the treatment group is lower than that of the control group.
- H1.b: The median probability of the predictions of the treatment group is closer to the true outcome of the event than that of the control group.
- H1.c: The average time spent on answering the survey is lower for the control group than for the treatment group.

4 Methods

4.1 Experimental design

The experiment was carried out using a Qualtrics survey entailing 10 real-life questions posed in a randomised order, with topics spacing from the Eurovision song contest to the Russian-Ukrainian war. The treatment presented the subjects with a market-like framework, while the control proposed a simple poll. In both cases the subjects needed to state the probability of a certain event resolving in a certain way, for example Ukraine winning the Eurovision song contest.

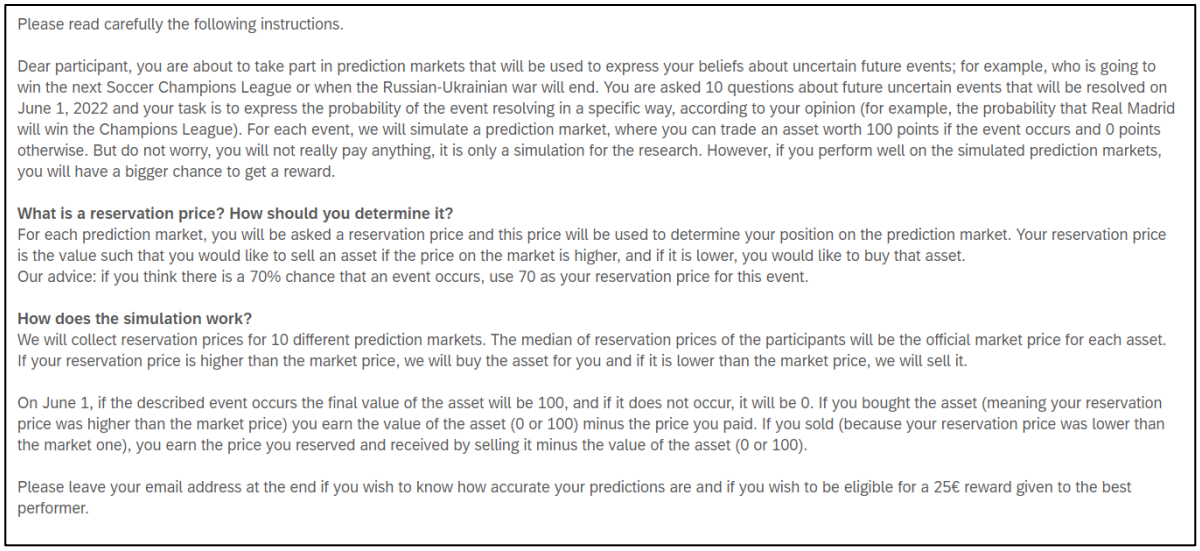
The experimental design passed the ethical check presented by the ethical questionnaire of Erasmus University, since it does not present any manipulation of the subjects except for a small monetary incentive (25€) and the treatment itself. The monetary incentive is used to encourage participation and boost the quality of the answers. It is rewarded both to the best performer of the treatment group and of the control group, if they decided to leave their email address. If they did not, the survey is completely anonymous and no data can be traced to any specific participant. The only ethical concern, since a payment is involved, is the participation of minors. Therefore, participants needed to be at least 18 years of age and any accidental minor participant is excluded from the results and is not eligible for any payment.

An *ex-ante* power analysis was performed to determine the estimated minimum sample size using the statistical software GPower (see Appendix for details). It determined that, to obtain a sample size sufficient to appreciate statistically significant effects of the treatment, at least 176 subjects must take part to the experiment, divided into treatment and control groups. The final number of subjects participating, after excluding incomplete responses and other non-suitable participants, was 181.

This thesis is a classical experimental design, with a treatment and a control group, carried out through a fully online survey, created with the software Qualtrics, using the licence kindly supplied by Erasmus Universiteit. Subjects, after agreeing to participate in the study and learning about the present incentives, are presented with the explanation of the task: in the case of the control group, subjects are asked to state the probability that they believe represents the actual chance of that event happening (for example, stating that they believe that Ukraine will win the Eurovision song contest with a 70% probability); in the case of the treatment group, subjects were also asked to state the probability representing the actual chance of that event happening, but bearing in mind the competitive element presented to them by the market framework. Specifically, the market condition

will be simulated in the following way: each subject will state his or her reservation price, in a range from 0 to 100, for a stock yielding 100 if the event resolves in the predicted way and 0 otherwise; after all respondents have answered, the median of the answers is calculated and computed as price of the stock; if the reservation price is higher than the market price, the subject will buy, obtaining a "long position" on the stock; if the reservation price is lower than the market price, the subject will sell, obtaining a "short position" on the stock; once the event is resolved, the value of the asset is determined (0 or 100); if the subject is long, he or she earns the value of the asset (0 or 100) minus the price they paid: if the subject is short, he or she earns the price they received minus the value of the asset (0 or 100). Therefore, subjects are incentivized to think more carefully about the answers in the treatment group. It is important to note that the monetary incentive is the same for both groups, as the best performer (most accurate) in the control group is rewarded with 25€ and the same applies to the treatment group. The total budget for the experiment, provided by the me, is 50€, as recommended by faculty guidelines.

For reference, this is what a subject placed in the treatment group is presented with, when reading the instructions of the experimental task:



Please read carefully the following instructions.

Dear participant, you are about to take part in prediction markets that will be used to express your beliefs about uncertain future events; for example, who is going to win the next Soccer Champions League or when the Russian-Ukrainian war will end. You are asked 10 questions about future uncertain events that will be resolved on June 1, 2022 and your task is to express the probability of the event resolving in a specific way, according to your opinion (for example, the probability that Real Madrid will win the Champions League). For each event, we will simulate a prediction market, where you can trade an asset worth 100 points if the event occurs and 0 points otherwise. But do not worry, you will not really pay anything, it is only a simulation for the research. However, if you perform well on the simulated prediction markets, you will have a bigger chance to get a reward.

What is a reservation price? How should you determine it?
For each prediction market, you will be asked a reservation price and this price will be used to determine your position on the prediction market. Your reservation price is the value such that you would like to sell an asset if the price on the market is higher, and if it is lower, you would like to buy that asset.
Our advice: if you think there is a 70% chance that an event occurs, use 70 as your reservation price for this event.

How does the simulation work?
We will collect reservation prices for 10 different prediction markets. The median of reservation prices of the participants will be the official market price for each asset. If your reservation price is higher than the market price, we will buy the asset for you and if it is lower than the market price, we will sell it.

On June 1, if the described event occurs the final value of the asset will be 100, and if it does not occur, it will be 0. If you bought the asset (meaning your reservation price was higher than the market price) you earn the value of the asset (0 or 100) minus the price you paid. If you sold (because your reservation price was lower than the market one), you earn the price you reserved and received by selling it minus the value of the asset (0 or 100).

Please leave your email address at the end if you wish to know how accurate your predictions are and if you wish to be eligible for a 25€ reward given to the best performer.

Figure 1: Instructions for the treatment group

4.2 Brier Score formula

The main variable tested in the experiment is the Brier score resulting from the predictions of the subjects. As explained during the literature review, the Brier score is a measure of accuracy of a probabilistic prediction of an event with a binary outcome. In the survey, the 10 questions were all formulated with binary resolutions (either the event is verified on June 1st, 2022, or it is not).

The math behind the Brier score is fairly simple:

$$Brier\ score = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

Where:

- N is the number of predictions under consideration
- t indexes the predictions from 1 to N
- f_t is the forecast, expressed as a probability from 0 to 1 for the t^{th} event

- o_t is the outcome (0 or 1) of the t^{th} event

As the reader may have noticed, the Brier score is also defined as the mean squared error between the predictions and the corresponding event outcomes. Larger differences between the probabilistic forecasts and event outcomes correspond to a greater error margin in the predictions. Therefore, a lower Brier score indicates greater accuracy. Moreover, since all squared errors lie between 0 and 1, the resulting scores are bound between 0 and 1. The main hypothesis implies that the Brier scores of the treatment group are, on average, lower than those of the control group. A secondary hypothesis suggests that the median distance of the predictions from the true outcome in the treatment group is lower than those of the control group. Additionally, a tertiary hypothesis states that the time spent answering the survey is, on average, higher for the treatment group, indicating a higher degree of focus. The main statistical test that is used is the standard t-test.

The design of the experiment is entirely between-subjects: even though the questions that subjects answer are the same in both groups, no subject is part of the control and of the treatment group at the same time, as the groups are mutually exclusive. Therefore, all variables of interest are entirely between-subjects.

4.3 Sample size

Before running the survey, it is important to run a power analysis, to estimate a sufficient minimum sample size to achieve a satisfactory result of the experiment. In this case, the open-source software GPower was used and default parameters were selected: a significance level of 0.05 and a power level of 0.80, both considered the gold standard for academic research. The estimated minimum sample size, calculated *a priori*, was found to be 176 subjects, to be equally allocated between control and treatment. The survey was distributed starting on May 3rd, 2022, and data collection was stopped on May 14th, 2022, resulting in 251 responses. After removing incomplete responses and one underage participant, the final sample size totals 181. 97 participants were presented with simple survey and 84 were served the treatment. As previously mentioned, underage participants were removed from the final sample due to ethical concerns given from the use of monetary incentives.

Randomization is applied to the experiment in two different ways: on one hand, the Qualtrics software was programmed to automatically assign every participant either to the control group or to the treatment group, making sure that they are equally and randomly distributed; on the other hand, to introduce further randomization and avoid biases from the order of the questions, the questions were presented to every participant in a completely randomized order.

4.4 Demographic characteristics of the sample

Before they were presented with the experimental tasks, participants were asked to give their demographic information. Specifically, the demographic characteristics of the final sample are as follows:

4.4.1 Age

The survey asked subjects to state their age through seven brackets, plus an option to express their will of not answering the question.

The minimum age stated was 18-24 years old and the maximum 55-64 years old. The mean age was 2.69, meaning that the majority of the subjects is between the second and the third bracket, slightly skewed towards the third one. The standard deviation (in terms of brackets) is 1.11 and the variance is 1.24.

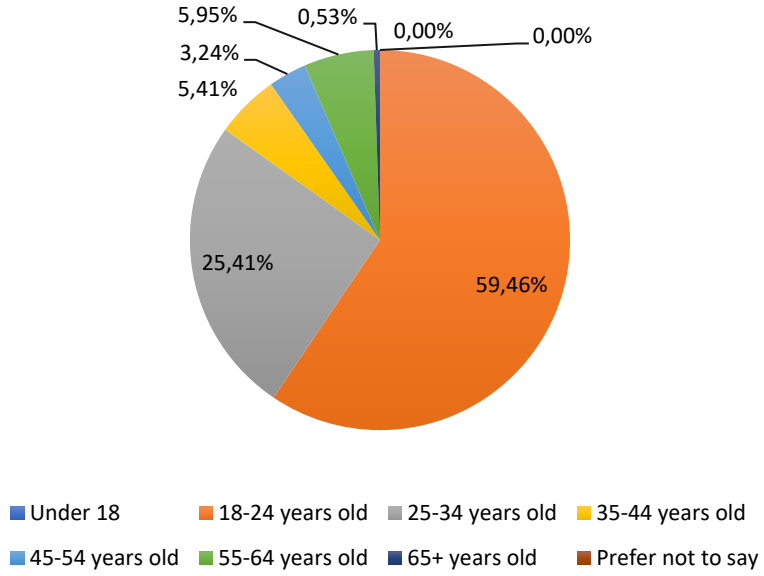


Figure 2: Pie chart of age of the participants

4.4.2 Gender

The survey asked subjects to state their gender as female, male, non-binary or self-describe. The option of not answering was also provided. The sample is quite balanced, as 54.59% of respondents declared themselves as female and 43.78% as male. The residual chose a different option but their weight is very marginal.

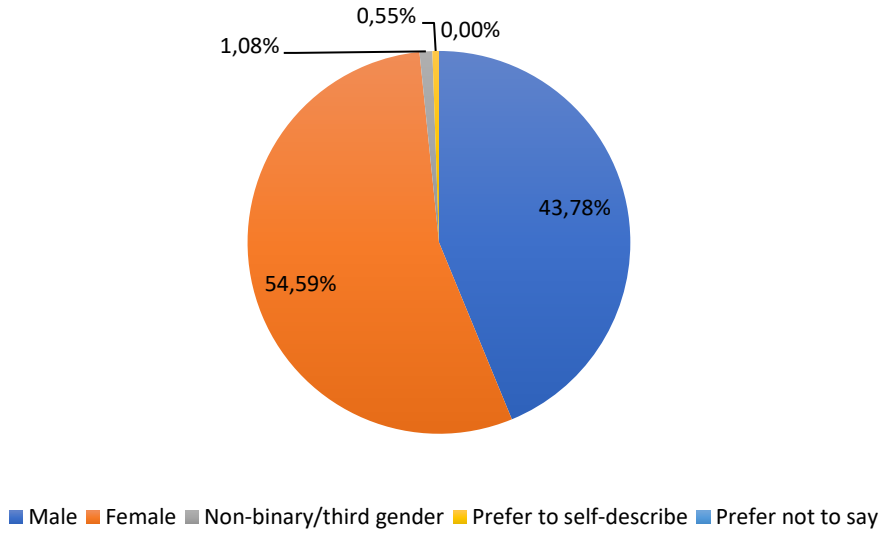


Figure 3: Pie chart of gender of the participants

4.4.3 Employment status

The survey asked subjects to state their current employment status by expressing a preference between six categories, which are as follows: Working full-time; Working part-time; Unemployed and looking for work; A Homemaker or stay-at-home parent; Student; Retired.

There were also two additional options, that gave the possibility of self-describing one's own profession or to avoid the question. The mean answer was 3.69, with the majority of the respondents being students or full-time workers. The standard deviation was 1.95 and the variance 3.80.

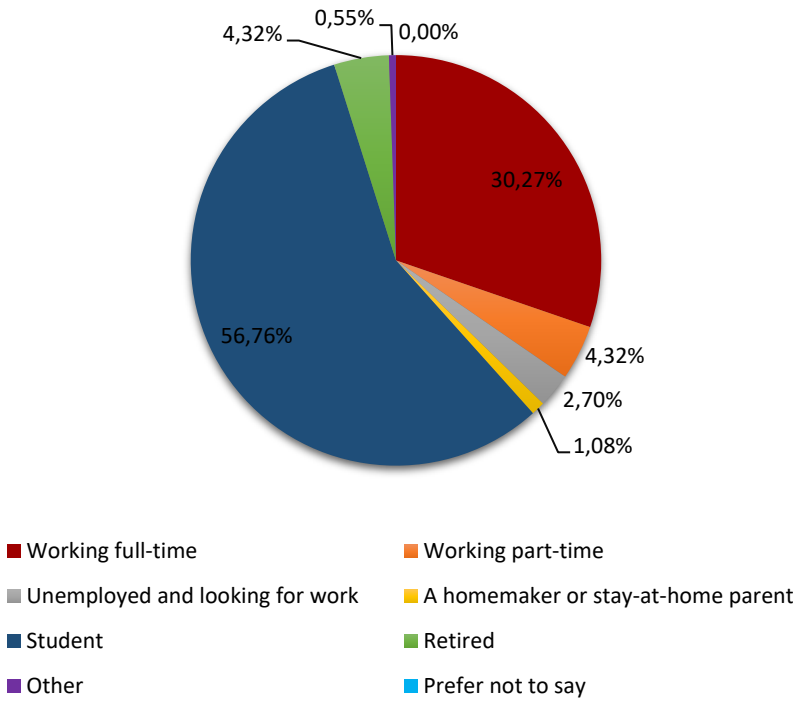


Figure 4: Pie chart of employment status of the participants

4.4.4 Nationality

Participants could choose between five default options (Dutch, Croatian, Italian, German, Spanish), which were based on the expected most common nationalities, and the option to simply state one's own nationality or avoid the question. The dominant nationalities are Italian and Croatian, followed by Dutch and Australian, as reflected by the mean of 2.88. The standard deviation was 1.32 and the variance 1.75.

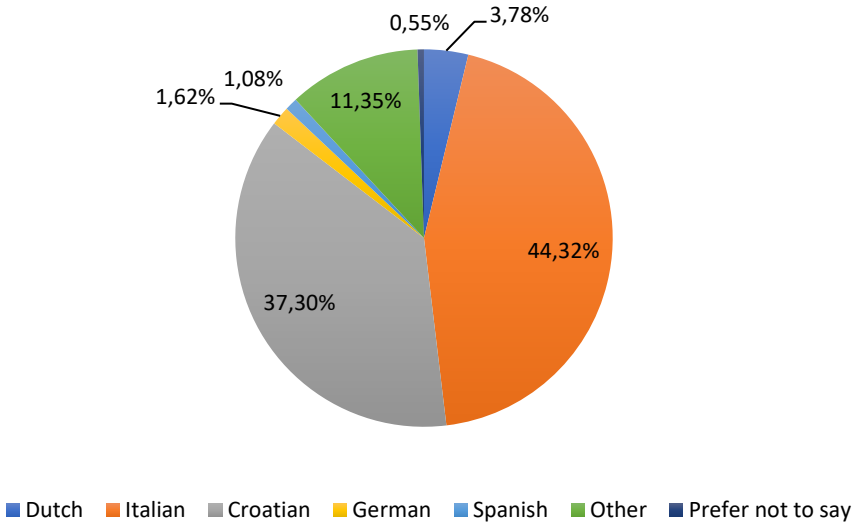


Figure 5: Pie chart of nationality of the participants

4.4.5 Education

The survey asked subjects to state their current level of education (last obtained). There was also an option to avoid the question. The mean was 2.88, reflecting the fact that the majority of the respondents has either completed only high school or has finished a Bachelor's degree. Another 20% is in possession of a Master's degree. The standard deviation was 1.32 and the variance was 1.75.

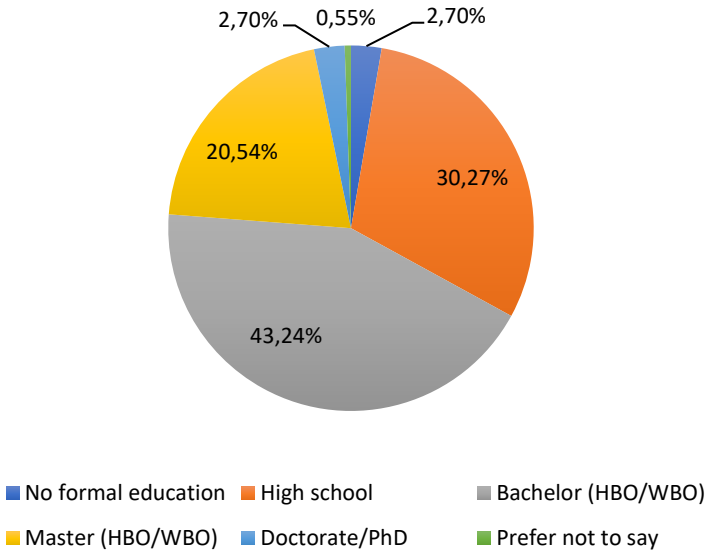


Figure 6: Pie chart of the education of the participants

4.4.6 Income

The survey asked subjects to state their personal yearly income through five brackets, with an additional option to express their will not to answer the question. The mean value was 2.18, reflecting the fact that more than half of the respondents declared less than €10,000 and a quarter between €10,000 and €39,999. The standard deviation was 1.74 and the variance 3.04.

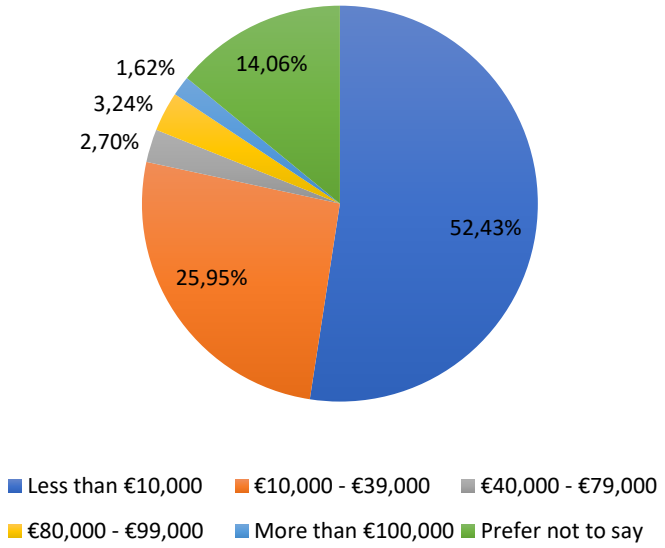


Figure 7: Pie chart of income of the participants

4.5 Experimental procedure

The survey took 5 to 7 minutes to complete, depending on which group the respondent is assigned to (the control group takes less time than the treatment) and on the subjects' speed of completion. The survey can be completed over a mobile device or a desktop computer and was optimised for both choices. There is no compensation for the time of invested by the subjects due to lack of resources, but there are competition and participation incentives: there is a prize of 25€ available for the best predicting performance in the control group and in the treatment group. A secondary incentive is to know how the respondent has ranked compared to others, element that was made stronger by the fact that groups of subjects knew each other (friends and relatives, mainly) and they engaged in friendly challenges. This was done to reflect a component of gamification of the experiment.

Subjects were recruited entirely through personal contact from the author, most commonly via WhatsApp and Telegram, followed by word of mouth between the subjects themselves, which were encouraged to share the survey link with people of their choice, which they most kindly did, usually with close friends and relatives. The survey link was also shared in some larger online groups, like the WhatsApp group of the Behavioural Economics students at the Erasmus School of Economics, as well as being accessible from the author's personal social media accounts. It was explained to every possible subject that the participation required a sufficient competence in the English language. This was the only other requirement that was announced, next to being at least 18 years of age.

Upon of opening the Qualtrics link, respondents from both the control and the treatment group were presented with a brief introduction, explaining that the survey is an experimental design for the Master Thesis. They are told that their answers are fully anonymous, except if they want to be eligible for the monetary reward or know the accuracy of their predictions afterwards. They are also informed

that they can withdraw their consent to use their data at any point, by contacting me by email, which was provided. Finally, they are notified that the estimated survey completion time is around 5 minutes. After answering demographical questions, subjects are randomly assigned by the software to either the control or the treatment group and are shown different instructions regarding their task.

The control group is shown a text explaining that the experiment they are participating in is about predicting the probability of a series of uncertain future events. Using an example question (not present among the actual predictions of the questionnaire), it is shown how to express their belief, namely by stating the probability that they believe represents the likelihood of the outcome in question. The text suggests that, if a subject believes that a certain event has a 70% probability of resolving in the stated way, they should express 70 when answering the question. It is also suggested to use extreme values (0 and 100) with caution. Finally, the subjects are again reminded of the optional possibility to leave their email at the end of the survey, to be eligible for the prize and to know the results. It is not possible to understand that this is a control group, as it is not stated in any part of the survey. The exact text is available in the appendix of this paper.

On the other hand, the treatment group is informed that they are about to take part in a prediction market, with a definition explaining what a prediction market is and for which purpose will be used (to predict uncertain future events). They are informed that they will bet on the outcome of the events through a reservation price but reassured that they do not need to use any real money, only virtual points. Subjects are notified that to every prediction corresponds an asset, which can be worth 100 points if the uncertain event resolves positively and 0 points if the event resolves negatively. It is then explained what a reservation price is, remarking its importance in the context of the prediction markets. Lastly, the market simulation is accurately described, specifying that subjects cannot know the market price when making the predictions (it will be calculated after, by taking the median of all reservation prices for each event). The transactions take place after the resolution of the events, in particular: if the reservation price of the subject is higher than the market price, they obtain a "long position" on the asset and the asset is automatically bought for them; if the reservation price of the subject is lower than the market price, they obtain a "short position" on the asset and the asset is automatically sold for them. After June 1st, 2022, each asset gains a value (either 0 or 100), depending on its resolution. Lastly, the individual performance of every subject is computed, following this rule: if the subject has a long position, they earn the value of the asset (0 or 100) minus the price they paid for the asset; if the subject has a short position, they earn the price they received minus the value of the asset. The exact text can be found in the appendix.

After the introduction and explanation of the tasks, both groups are presented with the same 10 questions, with topics spacing from the Russian-Ukrainian war to the Eurovision song contest. The questions are presented in a random order to each subject. In both groups subjects express their predicted probabilities through a slider, freely varying from 0 to 100. The complete list of questions can be found in the appendix.

The last part of the survey entails a brief conclusion, where I thank the participants for their time and dedication and remind them that they can leave, in a designated space, their email address to be eligible for the monetary reward granted to the best performers and to be informed about their results. The exact text can also be found in the appendix.

4.6 Analysis

As explained above, according to the roadmap of this work, data collection from the respondents of the survey was to be completed by the end of May and every event was supposed to be resolved by June 1st 2022. While data collection was completed as planned, the resolving of all events was partly

delayed by the persistence of tennis player Rafael Nadal, which went on to win the Roland-Garros on June 5th 2022. Including his win in the French Open, only two other forecasted events assumed a truth value (value 1 for the Brier score calculations and value 100 for the market simulation): a Ukrainian band won the Eurovision song contest in Torino and the value of the Amazon stock was above $1000 per stock on the designated day. All other events were not verified and assumed a truth value of 0.

According to the original plan and to the common practices that are used in experiments with a control and a treatment group, a balance test is performed to determine if the sample is balanced, meaning that the experimenter assesses whether the distribution of covariates is similar between the treatment and the control group. However, an influential strain of literature (Altman, 1985; Bruhn and McKenzie, 2008) argues that balance tests may be problematic, in economic research (and not only). As Altman notes, the theoretical foundation of these tests is unstable, because they amount to assessing the probability of something having occurred by chance after the experimenter knows that it did indeed occur by chance (no biases were willingly introduced). Moreover, these tests are only necessary if there is a reasonable worry that there was a problem in the randomisation procedure of the experiment. As this is not the case, it is far more probable that the use of a balance t-test would be improper and introduced errors to the results. Lastly, the small sample size is likely to sharpen these problems, rather than tame them. Therefore, as I am reasonably sure that the randomisation process was not faulty, no balance test was performed before continuing the analysis, as it could be more detrimental than beneficial. Consequentially, the sample is assumed to be balanced.

This analysis is set to confirm the main research hypothesis of this work:

H1: the accuracy of the predictions expressed by the treatment group is higher than those expressed by the control group.

This hypothesis is articulated in three sub-hypotheses, which represent three different ways to test the main hypothesis:

- H1.a: the average Brier score of the treatment group is lower than that of the control group.
- H1.b: the median probability of the predictions of the treatment group is closer to the true outcome of the event than that of the control group.
- H1.c: the average time spent on answering the survey is lower for the control group than for the treatment group.

Every sub-hypothesis is discussed in detail in the next section.

5 Results

The first steps that were taken when computing results were aimed at recontacting the subjects that left their emails and took part in the competition in the shortest time possible. Hence, the data was downloaded from Qualtrics and exported, without modifications, to a Microsoft Excel spreadsheet. Using Excel, the data was cleaned (as discussed in the previous section), eliminating unsuitable data points. Then, the Brier scores for every participant were calculated (see section 4.2 for details) separately for the control and treatment groups. The ranking that derives for the control group (the lower score, the higher accuracy in the predictions) was used to compute the winner for this group. On the other hand, while the Brier score of the treatment group was also computed, the ranking of the participants of this group was calculated following their “market result”, meaning the amount of points that they gained (or lost) in the market simulation (the more points, the better the result). The detailed description of how the market simulation worked can be found in section 4.1.

Once the rankings were computed and the winners of the two groups were selected, two emails, one per group, were sent to the subjects, announcing the winners. The emails were provided in the tree most common languages of the respondents, which are English, Italian and Croatian, to increase readability. Due to privacy concerns, the full rankings were not disclosed to the subjects (to protect the email addresses) but the participants were encouraged to reach out in case they wanted to know more details (for example, knowing their position inside their ranking) or provide feedback. I am happy to say that many of them did, showing a substantial interest in the matter. Both the winners received a bank transfer with the promised amount, 25€. The exact text of the emails announcing the winners can be found in the appendix. The complete rankings are, however, not present due to privacy concerns and are known only to the author. Some of the participants, around 20% of the total, demonstrated a deeper interest in the experiment and reached out, asking for more information. In this case, their questions were answered and their ranking was revealed to them, without revealing any data of other subjects (for example, "you finished 13th out of 87").

The following results describe the resulting Brier scores for the treatment and the control groups:

Treatment:

	Percentiles
1%	.03825
5%	.08357
10%	.11954
25%	.15826
50%	.20327
75%	.264695
90%	.31791
95%	.34975
99%	.53944

Obs	84
Mean	.2142329
Std. Dev.	.0873542
Variance	.0076308
Skewness	.8051779
Kurtosis	4.646901

Control:

	Percentiles
1%	.07695
5%	.12256
10%	.13292
25%	.16817
50%	.22359
75%	.26303
90%	.29901
95%	.35664
99%	.436

Obs	97
Mean	.2216609
Std. Dev.	.0713851
Variance	.0050958
Skewness	.3843062
Kurtosis	3.024319

5.1 H1.a

The first sub-hypothesis set out to test if the treatment produced more accurate results than the control, through a comparison of the average Brier scores of the two groups. As explained before, the Brier score is a measure of accuracy for predictions and can take a value between 0 and 1. The lower the Brier score, the more accurate the prediction. As the Brier score is bound between 0 and 1, the author argues that there is no need to remove extreme values and that the mean of the Brier score of the treatment and control groups are an efficient measure of average accuracy of the predictions, for each group respectively. This follows from the fact that there cannot be a high variance for a Brier score, due to the nature of score itself. A standard Student t-test was then done on the entire dataset of completed responses, testing the statistical significance of the difference between the average Brier scores of the treatment and control groups. In particular, the average Brier score of the control group was 0.222, while the average Brier score of the treatment group was 0.214. It can be seen that this partly supports the hypothesis, as the second score is slightly lower than the first one. However, the Student t-test, used to test the statistical significance of the difference between the means of the two groups, gives a negative result: with a p-value of p=.265 and the relevant statistic of t(179)=.624, we cannot reject the null hypothesis that the two means are not different at any relevant significance level (see appendix for details). To complement the analysis, a simple OLS regression with clustering at the individual level was also performed, to inquire if there is any correlation between the Brier scores and the fact that the subject was part of the treatment or control group. Even though there is a slight correlation in the expected sense (people in the control group have a higher Brier score, hence a lower accuracy), this correlation is not significant at any relevant significance level.

Therefore, even though the results seem promising at first, when statistically tested with an appropriate methodology they are revealed to be not significant. Sub-hypothesis H1.a is not supported.

5.2 H1.b

The second sub-hypothesis set out to test if the treatment produced more accurate results than the control, through a comparison of the absolute distance between the median predictions of each group (question by question) and the true *ex-post* value of that event. The use of median predictions value is justified by the need of testing the accuracy of the predictions of the treatment group in another way, to account for possible problems with the use of the Brier score or other measures. Specifically, this is how this sub-hypothesis was tested: firstly, the median value of every prediction (for every question) of both groups was computed; secondly, the absolute distance between the true outcome of every given event and the median predictions was measured; and finally, a paired Student t-test was run to check if the treatment group had a lower distance between the median predictions and the true outcomes (0 or 100). According to this test (see appendix for details), the average distance of the control group was -8.8 and the average distance for the treatment group was -4.5. The one-sided p-value of the t-test testing if the difference between the mean distances is <0 is p=.006 and a t statistic of t(9)=-3.124, meaning that we can reject the null hypothesis that there is no difference between the means with a 1% significance level. To complement the analysis, a simple OLS regression with clustering at the individual level was also performed, to inquire if there is any correlation between the mean distances and the fact that the subject was part of the treatment or control group. No significant correlation was found.

Since this difference is significant and is in the expected direction (meaning that the control group has a larger average distance between median predictions and true outcomes), we can say that sub-hypothesis H1.b is not rejected with a 1% significance level.

5.3 H1.c

The third sub-hypothesis set out to test the higher complexity of the treatment by comparing the average time that subjects took to complete the survey, both in the treatment and the control. Given that the time of completion is used as a proxy for perceived complexity and, consequently, effort exerted in the task. Some measures were taken to avoid obtaining excessively large values: firstly, the software used had an automatic system that marked responses as incomplete after a set amount of time; secondly, considering only the responses marked as completed, the data was cleaned of outliers, with the aid of a boxplot. On this final set of responses, the average duration of completion of the two groups was calculated. Respondents of the control group took, on average, 402 seconds to complete the survey. On the other hand, respondents of the treatment group took, on average, 479 seconds. This confirms the ex-ante hypothesis. A Student t-test was then performed using to calculate the statistical significance of the difference between the mean duration of the control and treatment groups. The one-sided p-value of the t-test testing if the difference between the means is >0 is p=.093 and the t statistic is $t(172.39)=1.235$, meaning that we can reject the null hypothesis that there is no difference between the means with a 10% significance level. Given these results, we can say that sub-hypothesis H1.c is not rejected, even though the significance is not very high.

As a complement the analysis, a simple OLS regression with clustering at the individual level was also performed, to inquire if there is any correlation between the time that the subjects took to complete the survey and the fact that the subject was part of the treatment or control group. Even though there is a correlation in the expected sense (people in the control group take less time to complete the survey), this correlation is not significant at any relevant significance level.

In summary, since sub-hypothesis H1.a is rejected, sub-hypothesis H1.b is not rejected at a 1% significance level and sub-hypothesis H1.c is not rejected at a 10% significance level, hypothesis H1 is only partially rejected, partly confirming that the treatment did enhance the accuracy of the predictions of the treatment group in comparison to the control one.

5.4 Correlation analysis

Given the relevant literature, which suggests that the predicting ability of an individual can be heavily influenced by several demographic characteristics, such as income, gender, nationality, education and others, I believe that a correlation analysis can be used to enquire if the results are systematically biased by any of the demographic characteristics asked in the survey (for example, if predictions made by Croatian women are, on average, more accurate than the mean of their group). In the survey, every respondent was presented a series of demographic questions, about their age, gender, employment status, nationality, level of education and level of income. To encourage more truthful answers, the questions did not require a punctual answer but asked to state a wider level, for example that the subject was between 25 and 34 years of age, therefore creating categorical data points. To test if any of these characteristics influenced the accuracy of the predictions, an OLS regression was performed, using as variable of interest the Brier scores and as covariates the demographic variables. In particular, since they are all categorical variables, the Brier scores from both groups were tested against each level of each categorical variable (see appendix for details). Moreover, the standard errors of the regression were clustered at the individual level, since there is more than one observation per subject. The vast majority of these correlations proves to be not significant at a 5% level, however there are two notable exceptions, both regarding the age of the respondents. Firstly, regarding the treatment group, there is a positive correlation between the Brier score and the age of the respondent, significant at a 5% level: the Brier score of a subject of 18-24 years old is, on average, 0.076 points higher than the base category (which was 25-34 years old, in this case), *ceteris paribus*. Secondly, there is a negative correlation between the Brier score and the income of the respondent, significant at a

5% level: the Brier score of a subject earning more than €100000 per year is, on average, 14.1 points higher than the base category (which was people earning less than €10000, in this case), *ceteris paribus*. No significant correlations are found for the control group.

These results suggests that the demographics of this sample mostly do not have a specific effect on the accuracy of the predictions (expressed through the Brier score). As these results are not very impactful (as they predict biases only from one category of a demographic variable) and, most importantly, do not point in the same direction, I conclude that the sample did not present significant correlations between the accuracy of predictions and the socio-economic status of the subjects, their nationality or gender. The only correlations found are a positive correlation between age and accuracy of predictions in the treatment group (the younger the subject, the higher their accuracy) and a positive correlation between income and accuracy of predictions (the wealthier the subject, the higher their accuracy).

6 Discussion

The main focus of this work is to study whether using a market framework can nudge people to have better results of predictions on future uncertain events. The predicted events covered a wide array of topics, from the stock markets to the Russian war in Ukraine, touching the Eurovision song contest.

The results of the testing of H1 partially confirm that the treatment improved the accuracy of the predictions of the subjects in this group. In particular, we infer this conclusion from the results of the three sub-hypotheses: the first one finds that subjects in the treatment group had a lower Brier score, on average, compared to the control group and therefore a higher average accuracy of the predictions, but this result is not statistically significant; the second sub-hypothesis, on the other hand, finds that the median predictions of the treatment group are closer to the true outcomes of the predicted events than those of the control group and this result is statistically significant; the third sub-hypothesis finds that, on average, subjects in the treatment group used more time to answer the survey, implying that they also employed more effort than the control group, and this result is statistically significant, even though at a lower significance level. Finally, the results of the additional correlation study confirm that there is no systematic bias in the sample regarding any of the demographic covariates of interest.

Although the results of H1 are not very strong, they are in line with what was expected by the author before the start of this work. In particular, they are in line with the stream of literature on prediction markets (see section 2.3), which argues that the market mechanism, when applied to predictions, can yield better results than other more traditional methods.

However, it must be noted that this work has several limitations, mostly due to the limited resources that I could employ, both logistically, economically and temporally. In most of the studies on this topic that preceded this one, the market element is implemented through several live interactions, often during the course of months, if not years. Since this is only a Master's thesis, both time and means are limited. The employed solution managed to simulate the market mechanism through a one-time interaction bet system. Naturally, this has limitations, the most prominent being the absence of actual interaction between the subjects and the fact they only simulated using money but none was actually used. As economic incentives for participation and better elicitation of effort are very important, this is undeniably a big limitation. However, since the survey was run mostly between friends and family of the author, it was somewhat compensated, given that the economic incentives were still present in the form of a prize to the best performers and that the subjects were intrinsically motivated by personal affection. A strong point of the sample is also its diversity: given my personal circumstances, it was easy to gather participants from a large number of European countries and different contexts (economically, culturally and socially), therefore accounting for possible biases that would affect a more homogeneous sample. A very diverse sample can help with avoiding biases in the data due to excessive income, nationality or educational homogeneity and help with the overall balancing of the survey. While I am satisfied by the number of responses gathered (as they are sufficient according to the *ex-ante* power analysis), the total final number of responses is not enough to find strong statistical significance for the tested hypothesis, which could have probably been found with more participants. On the other hand, the fact that the market was not live, while it does not allow for a complete analysis of a prediction market, it focuses on the competitive element of said markets and its effects on intrinsically motivated forecasters, which is surely an interesting element partly overlooked in past research.

Overall, this study certainly has major limitations, mainly derived from lacking means and time constraints, but it manages to underline that even using only a very partial market framework focusing on remote competition between participants prompts them to better forecast the outcomes of future

uncertain events. It would be ideal to expand on this research, possibly using stronger economic incentives and giving subjects money to bet with, which they could keep if they perform well, as well as including in the study a larger number of subjects and more ways to test the accuracy of their predictions.

7 Conclusion

Prediction markets have been proposed as an innovative tool for predicting future events and can already be found on the internet both for research, like the Iowa Electronic Markets, and commercial purposes, like TradeSports and NewsFutures's World News Exchange, which organise sports, financial and political betting markets. The potential for future employment in the private and public sector is immeasurable, as prediction markets can effectively function as decision support systems. In this thesis, there was an attempt to set up a non-live simulation of a betting market on uncertain future events and compare its results with a control group in which predictions on the same questions were elicited through simple polling.

While this work shows the potential that prediction markets can bring to increase the accuracy of predictions of future uncertain events, further research is needed. In particular, one of the flaws that were underlined about this methodology is that, to achieve significant results, a large number of participants (at least 100) is required. Even though this is maybe not a problem for a hypothetical use as decision support when developing public policies, it is easy to envision logistical and economical problems when creating prediction markets supporting decision for private companies or divisions within companies. A possible solution, which is already being explored, is the use of online aggregators capable of putting together thousands of participants at an affordable cost. Therefore, I suggest replicating this study with a larger number of participants, to confirm or disprove its main findings, as well as exploring new methodologies to reduce the number of participants needed for a prediction market to be effective, allowing for the creation of small teams of superforecasters that would participate in prediction markets supporting decision makers, in both public and private sectors.

8 References

Altman, D. G. (1985). Comparability of randomised groups. Journal of the Royal Statistical Society: Series D (The Statistician), 34(1), 125-136.

Berg, J. E., & Rietz, T. A. (2003). Prediction markets as decision support systems. Information systems frontiers, 5(1), 79-93.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly weather review, 78(1), 1-3.

Bruhn, M., & McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. American economic journal: applied economics, 1(4), 200-232.

Casati, B., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A., Pocernich, M., ... & Mason, S. (2008). Forecast verification: current status and future directions. Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling, 15(1), 3-18.

Danz, D., Vesterlund, L., & Wilson, A. J. (2020). Belief elicitation: Limiting truth telling with information on incentives (No. w27327). National Bureau of Economic Research.

Falk, A., & Szech, N. (2013). The Systematic Place of Morals in Markets—Response. Science, 341(6147), 714-714.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. Cognitive psychology, 18(3), 253-292.

Elliott, G., & Timmermann, A. (2008). Economic forecasting. Journal of Economic Literature, 46(1), 3-56.

Hayek, F. A. (1945). The use of knowledge in society. The American economic review, 35(4), 519-530.

Hankins, R., & Lee, A. (2011). Crowd sourcing and prediction markets. In CHI (Vol. 11, pp. 17-20).

Hogarth, R. M., & Makridakis, S. (1981). Forecasting and planning: An evaluation. Management science, 27(2), 115-138.

Hossain, T., & Okui, R. (2013). The binarized scoring rule. Review of Economic Studies, 80(3), 984-1001.

Katsagounos, I., Thomakos, D. D., Litsiou, K., & Nikolopoulos, K. (2021). Superforecasting reality check: Evidence from a small pool of experts and expedited identification. European journal of operational research, 289(1), 107-117.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. Journal of Experimental Psychology: Human learning and memory, 6(2), 107.

Maciejovsky, B., & Budescu, D. V. (2020). Too much trust in group decisions: Uncovering hidden profiles by groups and markets. Organization Science, 31(6), 1497-1514.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., ... & Tetlock, P. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. Perspectives on Psychological Science, 10(3), 267-281.

Murphy, A. H. (1997). Forecast verification. Economic value of weather and climate forecasts, 19, 74.

Schoemaker, P. J., & Tetlock, P. E. (2016). Superforecasting: How to upgrade your company's judgment. Harvard Business Review, 94(5), 73-78.

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1), 37-45.

Trautmann, S. T., & van de Kuilen, G. (2015). Belief elicitation: A horse race among truth serums. The Economic Journal, 125(589), 2116-2135.

Tziralis, G., & Tatsiopoulos, I. (2007). Prediction markets: An extended literature review. The journal of prediction markets, 1(1), 75-91.

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. Journal of economic perspectives, 18(2), 107-126.

9 Appendix

9.1 Appendix A - Instructions and preparations

GPower calculation:

t tests - Means: Difference between two independent means (two groups)

Analysis: A priori: Compute required sample size

Input: Tail(s) = One

Effect size d = 0.5

α err prob = 0.05

Power (1-β err prob) = 0.95

Allocation ratio N2/N1 = 1

Output: Noncentrality parameter δ = 3.3166248

Critical t = 1.6536580

Df = 174

Sample size group 1 = 88

Sample size group 2 = 88

Total sample size = 176

Actual power = 0.9514254

Introduction (common to both groups)

Welcome and thank you for partaking in this study!

I am a student from the Erasmus University Rotterdam and this survey is the main part of my Master's Thesis in Behavioral Economic. I kindly ask that you answer truthfully and to the best of your abilities. It is important to mention that your participation in this study is fully anonymous. This means that no information is collected that could identify you and thus we cannot attribute any data to you, except for the possibility of obtaining a reward for participating. At the end of the survey there is the option of leaving your personal email, to have the chance of winning a monetary reward (the best performer will receive 25€). This survey is also voluntary and you can withdraw your consent at any time. If you wish to withdraw at a later stage or you would like to know more about our study and its results, you can contact me via this email: 619342em@student.eur.nl.

The survey will take you about 5 minutes to complete.

Control group instructions:

Please read the following instructions carefully.

Polling

Dear participant, you are about to take part of a simple poll on the probability that a series of uncertain future events will resolve in a specific way, for example who is going to win the next Soccer Champions League or when the Russian-Ukrainian war will end. You are asked 10 questions about future uncertain events that will be resolved on June 1, 2022 and you need to express the probability of the event resolving in a specific way, according to your opinion (for example, the probability that Real Madrid will win the Champions League). You will directly state the probability (between 0 and 100) that represents your belief of that event occurring. For example, if you are reasonably sure that Real Madrid will be the next winner of the Champions League, you should state a high enough probability (maybe 80%). Please use the extreme values (0 and 100) only if you are extremely sure about your prediction.

Please leave your email address at the end if you wish to know how accurate your predictions are and if you wish to be eligible for a 25€ reward given to the best performer.

Treatment group instructions:

Please read carefully the following instructions.

Dear participant, you are about to take part in prediction markets that will be used to express your beliefs about uncertain future events; for example, who is going to win the next Soccer Champions League or when the Russian-Ukrainian war will end. You are asked 10 questions about future uncertain events that will be resolved on June 1, 2022 and your task is to express the probability of the event resolving in a specific way, according to your opinion (for example, the probability that Real Madrid will win the Champions League). For each event, we will simulate a prediction market, where you can trade an asset worth 100 points if the event occurs and 0 points otherwise. But do not worry, you will not really pay anything, it is only a simulation for the research. However, if you perform well on the simulated prediction markets, you will have a bigger chance to get a reward.

What is a reservation price? How should you determine it?

For each prediction market, you will be asked a reservation price and this price will be used to determine your position on the prediction market. Your reservation price is the value such that you would like to sell an asset if the price on the market is higher, and if it is lower, you would like to buy that asset.

Our advice: if you think there is a 70% chance that an event occurs, use 70 as your reservation price for this event.

How does the simulation work?

We will collect reservation prices for 10 different prediction markets. The median of reservation prices of the participants will be the official market price for each asset. If your reservation price is higher than the market price, we will buy the asset for you and if it is lower than the market price, we will sell it.

On June 1, if the described event occurs the final value of the asset will be 100, and if it does not occur, it will be 0. If you bought the asset (meaning your reservation price was higher than the market price) you earn the value of the asset (0 or 100) minus the price you paid. If you sold (because your reservation price was lower than the market one), you earn the price you reserved and received by selling it minus the value of the asset (0 or 100).

Please leave your email address at the end if you wish to know how accurate your predictions are and if you wish to be eligible for a 25€ reward given to the best performer.

Conclusion of survey
Thank you for your effort and dedication. There is a 25€ reward for the top performer in this experiment. If you wish to be eligible for it or just know how accurate your predictions were, please leave an email address in the box below. You will be contacted in June, when the data will be analysed.

This is entirely optional.

List of questions for survey

1. Russia will invade Moldova before June 1.
2. Kharkiv will be under Russian control on June 1.
3. At least one NATO country will invoke Article 5 of the North Atlantic Treaty by June 1, 2022.
4. Ukraine will win the Eurovision Song Contest 2022.
5. Italy will win the Eurovision Song Contest 2022.
6. Rafael Nadal will win the French Open 2022 (Roland-Garros).
7. Bosnia and Herzegovina will fall apart before June 1, 2022.
8. SpaceX's Starship will reach orbit before June 1, 2022.
9. TESLA will be worth $1000 per stock or more on the Wall Street stock exchange by June 1, 2022.
10. AMAZON will be worth $1000 per stock or more on the Wall Street stock exchange by June 1, 2022.

Results email control group

The results of your predictions are in! / I risultati delle vostre predizioni sono pronti! / Rezultati Vaših predviđanja su tu!

Dear participants,

Every event on which you made predictions has finally wrapped up and I made the necessary calculations and analyses, creating a ranking. First of all, let me thank you all from the bottom of my heart for your time and dedication in helping me in this research, you were wonderful. Now that the experiment is over, I can tell you more details on how it worked: when you clicked on the link and started the survey, you were randomly assigned to either to the “control” group (the simple polling)

or the “treatment group”, the market simulation. There was an equal chance of getting either group and the software was automatically balancing the number of participants in each group. The main point of the experiment was to see if participants in the market simulation would be, on average, more accurate in their predictions than participants in the control group. I am happy to tell you that this was the case. There are two separate rankings, one for the control and one for the treatment group. You were part of the control group. These results were compiled using the Brier Score, which means that the lower is your score, the better were your predictions.

Congratulations to matte.oldani99@gmail.com for the top score! Please respond to this email to claim your 25€ prize.

If you are interested in learning more details about this experiment or have any comments or questions, feel free to email me as well.

Thank you again for participating.

Kind regards,

Eugenio Magni

//

Care partecipanti e cari partecipanti,

Tutti gli eventi su cui avete espresso predizioni si sono concretizzati e io ho completato i necessari calcoli ed analisi, compilando una classifica. Prima di tutto, vi ringrazio tutti dal profondo del cuore per il vostro tempo e la dedizione che avete riservato per aiutarmi nella mia ricerca, siete stati fanctastici. Ora che l’esperimento è concluso, posso rivelarvi più dettagli sul suo funzionamento: quando avete cliccato sul link ed iniziato il sondaggio, siete stati automaticamente e randomicamente assegnati dal software o al gruppo di “controllo” (il sondaggio semplice) o al gruppo di “trattamento”, la simulazione di mercato. C’era una equa possibilità di essere in uno o l’altro gruppo ed il software ha automaticamente bilanciato il numero di partecipanti in ogni gruppo. Lo scopo principale dell’esperimento era capire se i partecipanti della simulazione di mercato sarebbero stati, in media, più precisi nelle predizioni dei partecipanti del gruppo di controllo. Sono felice di dirvi che ciò si è verificato. Ci sono due classifiche separate, una per il gruppo di controllo ed una per il gruppo di trattamento. Tu eri parte del gruppo di controllo. I risultati sono stati computati usando il Brier score, il che significa che più è basso lo score, maggiore è stata l’accuratezza.

Congratulazioni a matte.oldani99@gmail.com per aver ottenuto il punteggio migliore! Per favore rispondi a questa mail per reclamare il tuo premio di 25€.

Se qualcuno è interessato a conoscere il proprio piazzamento in classifica o qualsiasi altro dettaglio sull’esperimento o ha domande o suggerimenti può scrivermi a questa mail.

Grazie ancora per la vostra gentile partecipazione.

Cordiali saluti,

Eugenio Magni

//

Dragi sudionici,

Svaki događaj za koji ste iznijeli svoja predviđanja konačno je zaključen, izvršeni su potrebni izračuni i analize te kreirana rang lista sudionika.

Prije svega, dopustite mi da vam svima od srca zahvalim na uloženom vremenu i trudu kojim se pomogli provesti ovo istraživanje. Sada, kada je eksperiment gotov, mogu vam reći više pojedinosti o tome kako je oblikovan: kada ste kliknuli na vezu i pokrenuli anketu, nasumično ste dodijeljeni ili u kontrolnu grupu (jednostavno ispitivanje) ili u ispitivanu grupu koja je simulirala tržište. Postojala je jednaka šansa za dobivanje bilo koje skupine, a softver je automatski balansirao broj sudionika u njima. Glavni cilj eksperimenta bio je pratiti hoće li sudionici u simulaciji tržišta u prosjeku biti točniji u svojim predviđanjima od sudionika u kontrolnoj skupini, što se takvim i pokazalo.

Postoje dvije odvojene rang liste sudionika, jedna za kontrolnu i jedna za ispitivanu grupu. Bili ste dio kontrolne grupe čiji su rezultati sastavljeni pomoću Brier Score, odnosno niži rezultat označava točnija predviđanja.

Čestitam matte.oldani99@gmail.com za najbolji rezultat! Odgovorite na ovu poruku e-pošte kako biste preuzeli svoju nagradu od 25 €.

Ako ste zainteresirani za više pojedinosti o ovom eksperimentu ili imate bilo kakve komentare ili pitanja, slobodno mi pošaljite poruku e-pošte.

Još jednom hvala što ste bili dio ovog istraživanja!

Lijepi pozdrav,

Eugenio Magni

Results email treatment group

The results of your predictions are in! / I risultati delle vostre predizioni sono pronti! / Rezultati Vaših predviđanja su tu!

Dear participants,

Every event on which you made predictions has finally wrapped up and I made the necessary calculations and analyses, creating a ranking. First of all, let me thank you all from the bottom of my heart for your time and dedication in helping me in this research, you were wonderful. Now that the experiment is over, I can tell you more details on how it worked: when you clicked on the link and started the survey, you were randomly assigned to either to the “control” group (the simple polling) or the “treatment group”, the market simulation. There was an equal chance of getting either group and the software was automatically balancing the number of participants in each group. The main point of the experiment was to see if participants in the market simulation would be, on average, more accurate than participants in the control group. I am happy to tell you that this was the case. Anyway, here are the results, detailing your level of accuracy. There are two separate rankings, one for the control and one for the treatment group. You were part of the treatment group. Since it was a market simulation, the ranking is based on how well you performed on the market, namely how many points you gained (or lost) during the experiment, based on your answers.

Congratulations to nicoguglie@hotmail.it for the top score! Please respond to this email to claim your 25€ prize.

If you are interested in learning more details about this experiment or have any comments or questions, feel free to email me as well.

Thank you again for participating.

Kind regards,

Eugenio Magni

Care partecipanti e cari partecipanti,

Tutti gli eventi su cui avete espresso predizioni si sono concretizzati e io ho completato i necessari calcoli ed analisi, compilando una classifica. Prima di tutto, vi ringrazio tutti dal profondo del cuore per il vostro tempo e la dedizione che avete riservato per aiutarmi nella mia ricerca, siete stati fanctastici. Ora che l'esperimento è concluso, posso rivelarvi più dettagli sul suo funzionamento: quando avete cliccato sul link ed iniziato il sondaggio, siete stati automaticamente e randomicamente assegnati dal software o al gruppo di "controllo" (il sondaggio semplice) o al gruppo di "trattamento", la simulazione di mercato. C'era una equa possibilità di essere in uno o l'altro gruppo ed il software ha automaticamente bilanciato il numero di partecipanti in ogni gruppo. Lo scopo principale dell'esperimento era capire se i partecipanti della simulazione di mercato sarebbero stati, in media, più precisi nelle predizioni dei partecipanti del gruppo di controllo. Sono felice di dirvi che ciò si è verificato. Ci sono due classifiche separate, una per il gruppo di controllo ed una per il gruppo di trattamento. Tu eri parte del gruppo di trattamento. Dato che era una simulazione di mercato, la classifica è basata sulla performance nel mercato, cioè su quanti punti

Congratulazioni a nicoguglie@hotmail.it per aver ottenuto il punteggio migliore! Per favore rispondi a questa mail per reclamare il tuo premio di 25€.

Se qualcuno è interessato a conoscere il proprio piazzamento in classifica o qualsiasi altro dettaglio sull'esperimento o ha domande o suggerimenti può scrivermi a questa mail.

Grazie ancora per la vostra gentile partecipazione.

Cordiali saluti,

Eugenio Magni

//

Dragi sudionici,

Svaki događaj za koji ste iznijeli svoja predviđanja konačno je zaključen, izvršeni su potrebni izračuni i analize te kreirana rang lista sudionika.

Prije svega, dopustite mi da vam svima od srca zahvalim na uloženom vremenu i trudu kojim se pomogli provesti ovo istraživanje. Sada, kada je eksperiment gotov, mogu vam reći više pojedinosti o tome kako je oblikovan: kada ste kliknuli na vezu i pokrenuli anketu, nasumično ste dodijeljeni ili u kontrolnu grupu (jednostavno ispitivanje) ili u ispitivanu grupu koja je simulirala tržište. Postojala je jednaka šansa za dobivanje bilo koje skupine, a softver je automatski balansirao broj sudionika u njima. Glavni cilj eksperimenta bio je pratiti hoće li sudionici u simulaciji tržišta u prosjeku biti točniji u svojim predviđanjima od sudionika u kontrolnoj skupini, što se takvim i pokazalo.

Postoje dvije odvojene rang liste sudionika, jedna za kontrolnu i jedna za ispitivanu grupu. Bili ste dio ispitivane grupe čija je rang lista temeljena na tržištu, odnosno koliko ste bodova na temelju danih odgovora dobili (ili izgubili) tijekom eksperimenta.

Čestitam nicoguglie@hotmail.it za najbolji rezultat! Odgovorite na ovu poruku e-pošte kako biste preuzeli svoju nagradu od 25 €.

Ako ste zainteresirani za više pojedinosti o ovom eksperimentu ili imate bilo kakve komentare ili pitanja, slobodno mi pošaljite poruku e-pošte.

Još jednom hvala što ste bili dio ovog istraživanja!

Lijepi pozdrav,

Eugenio Magni

9.2 Appendix B – Statistical results

Details H1.a

```
----------------------------
                       (1)
             Brierscore~e
----------------------------
control           0.00743
                  (0.536)

_cons               0.214***
                  (0.000)
----------------------------
N                     181
R-sq                0.002
adj. R-sq          -0.003
----------------------------
p-values in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

Details H1.b

```
----------------------------
                       (1)
             Ab. differences
----------------------------
control           0.00872
                  (0.679)

_cons               0.341***
                  (0.000)
----------------------------
N                     181
R-sq                0.004
adj. R-sq          -0.002
----------------------------
p-values in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

Details H1.c

	(1)
	Durationinseconds
control	-77.00
	(0.187)
_cons	479.9***
	(0.000)
N	181
R-sq	0.010
adj. R-sq	0.004

p-values in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Correlation analysis

	(1)	(2)
	Brier_control	Brier_treatment
_Iagecat_2	0.0329	-0.0795**
	(0.154)	(0.004)
_Iagecat_3	0.101	-0.00276
	(0.078)	(0.943)
_Iagecat_4	0.0244	-0.0957*
	(0.657)	(0.028)
_Iagecat_5	0.0103	-0.0596
	(0.748)	(0.269)
_Igenderca~2	0.0135	0.0387*
	(0.461)	(0.038)
_Igenderca~3	0	0.0927
	(.)	(0.202)
_Ijobcat_2	-0.0382	-0.0683
	(0.371)	(0.129)
_Ijobcat_3	0.0618	0.0229
	(0.157)	(0.773)
_Ijobcat_4	-0.00863	0
	(0.854)	(.)
_Ijobcat_5	0.0224	-0.0928*
	(0.505)	(0.011)
_Ijobcat_6	-0.0235	-0.0901
	(0.645)	(0.092)
_Inational~2	-0.00696	0.0519

	(0.870)	(0.283)
_Inational~3	-0.0305	0.00163
	(0.503)	(0.975)
_Inational~4	-0.0387	0
	(0.520)	(.)
_Inational~5	0.0402	0
	(0.399)	(.)
_Inational~6	-0.0157	0.0431
	(0.749)	(0.429)
_Ieducatio~2	-0.0278	0.0273
	(0.510)	(0.607)
_Ieducatio~3	-0.0659	0.0162
	(0.078)	(0.760)
_Ieducatio~4	-0.00745	-0.00968
	(0.820)	(0.856)
_Ieducatio~5	-0.00725	0.0507
	(0.872)	(0.298)
_Iincomeca~2	0.00476	-0.0419
	(0.801)	(0.102)
_Iincomeca~3	-0.0606	0.00153
	(0.114)	(0.955)
_Iincomeca~4	0.00625	-0.0738*
	(0.824)	(0.039)
_Iincomeca~5	-0.0627	-0.141**
	(0.333)	(0.009)
_cons	0.243***	0.247**
	(0.000)	(0.003)
N	97	84
R-sq	0.210	0.339
adj. R-sq	-0.039	0.115

p-values in parentheses
* p<0.05, ** p<0.01, *** p<0.001