ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis Behavioural Economics

Truth-Telling Incentives for Unverifiable Questions Using Choice-Matching

Student: Alejandro Arienti
Student number: 530042

Supervisor: Prof. Dr. Han Bleichrodt

Second assessor: Dr. Benjamin Tereick

Date final version: July 20th, 2020

Abstract

What happens when you wish to elicit people's true preferences, beliefs, or behaviours, but have no way of verifying if these are indeed truthful? As a social species, there are many reasons why humans might (un-)consciously choose to conceal their preferences, beliefs, or behaviours. However, new methods have been developed to incentivize truth-telling. This master thesis tests one of these methods developed by Cvitanić, Prelec, Riley, and Tereick (2019): choice-matching. The investigation builds upon research performed by Baillon, Bleichrodt, and Granic (2020) who tested the effectiveness of Drazen Prelec's (2004) Bayesian Truth Serum on questions about subjective well-being and health. They found a very limited effect of incentives. By conducting a survey, this investigation explores the effectiveness of truth-telling incentives using choice-matching on questions about subjective well-being, health, and behaviours & beliefs during the coronavirus COVID-19 pandemic. In line with Baillon et al. (2020) findings, the results suggest that truth-telling incentives do not have an effect on questions about health and subjective well-being, as long as no other cognitive biases are present. However, the results suggest that choice-matching incentives might encourage respondents to confess to socially undesired misconducts.

Key words: *choice-matching, truth-telling, incentives, Bayesian Truth Serum, subjective well-being, health, behaviour*

Acknowledgements

I would like to express my great appreciation to Prof. Dr. Han Bleichrodt for his valuable guidance and constructive feedback for writing this master thesis. I would also like to thank him for always being available and helping me finish my research in time.

I would also like to thank my friends and family for their support.

Finally, I would like to thank my wife for her patience, love, and encouragement during this whole process.

Table of Contents

Introduction

Researchers are interested in studying people's choices, preferences, beliefs, and behaviours. However, what happens when people consciously (or unconsciously) choose to behave differently or to report untruthfully? The reasons for choosing to act dishonestly are many. It might be unintentionally, because of having a bad memory for example (Hyman, 1944). Or it might be because people are embarrassed of their beliefs or preferences for some reason and choose to hide these (Edwards, 1953; Stanley Warner, 1965; Preisendörfer & Wolter, 2014; Adam Berinsky, 2018). Different approaches have been developed to tackle this issue. Some of these rely on behavioural biases and survey design, such as randomization, framing, default biasing, among others. Other more recent truth-telling mechanisms have also been developed. One of the most innovative ones is the Bayesian Truth-Serum (BTS) developed by Drazen Prelec (2004). Weaver & Prelec (2013) used the BTS method with financial incentives to test whether they could induce truth-telling from respondents. They succeeded. However, the BTS methodology is quite complex and hence difficult to comprehend by the general population. This, unfortunately, may weaken the BTS's power. If respondents do not understand why truth-telling is in their best interest, then they may become immune to the effects of BTS. Cvitanić, Prelec, Riley, and Tereick (2019) have come up with a solution to this problem which they named *choice-matching*. Under this newer truth-telling mechanism, participants play a forecasting game and get paid according to how good their predictions are. The underlying methodology of *choice-matching* causes participants to unconsciously respond honestly as they are trying to "compete" in the game. Explaining to participants the instructions of a game is much simpler than explaining the reasons why BTS might work and trying to convince them that it is in their best interests to respond truthfully. Therefore, *choice-matching* seems to be a promising and innovative truth-telling mechanism.

The intention of this research is to test the effectiveness of *choice-matching* on questions for which the survey conductor has no way to verify that respondents are being honest, and which may include a hypothetical bias. Furthermore, this research builds upon an investigation performed by Baillon, Bleichrodt, and Granic (2020) who tested the effect of incentives using BTS on subjective well-being and health-related questions. Taking this into consideration leads to the following research question:

Do truth-telling incentives have an effect on questions about subjective well-being, health, and beliefs & behaviour?

In order to answer the research question, this master thesis is structured as follows: Chapter One will focus on the Literature Review and findings from previous research. Chapter Two will discuss the methodology, explaining in more detail the *choice-matching* mechanism, the data collection, and data description. Chapter Three deals with the analysis and discussion of the results, as well as with the limitations to this investigation. The conclusion is included in Chapter Four.

Chapter One: Literature Review

There have been numerous attempts to incentivize people to tell the truth in situations where it is not possible to know whether a person is being truthful or not. In particular, this has been a challenging endeavour in opinion research, where the survey conductor has no way to prove that the respondents' answers reflect their true opinions, beliefs, behaviours, or preferences (Weaver & Prelec, 2013). In 1944, Herbert Hyman investigated survey responses from the Division of Surveys of the Office of War Information's Bureau of Special Services in the United States of America (Hyman, 1944). In his investigation, Herbert showed that indeed some people did not respond truthfully to polling questions about highly socially desirable behaviour (redemption of war bonds, display of government advertising, and worker absenteeism) (Hyman, 1944). In his conclusion, Herbert stated that "the results of public opinion polls should be used only with the greatest caution" (Hyman, 1944, p.559). Parry and Crossley (1950) questioned the validity of surveys in the field of public opinion and market research, arguing that the conditions under which a respondent expresses his or her opinion might differ significantly from the conditions in the real world.

There are different explanations as to why people might not respond as truthfully as the survey conductor would wish. Herbert Hyman (1944) argues that this mismatch might be due to errors of recall; that is, some people might just have a bad memory. Preisendörfer and Wolter (2014) investigated whether people would disclose having been convicted of a criminal offence. In their research, the authors argued that respondents underreport behaviours that are seen as socially undesirable (Preisendörfer & Wolter, 2014). Their findings are in line with Edwards (1953) who mentions that participants might choose to be untruthful in order to hide socially undesirable behaviour or beliefs. In politics, for example, Adam Berinsky (2018) investigates whether respondents are truthful when they claim to be in favour (or against) a certain political rumour, arguing that they may not believe a particular rumour to be true, but respond as if they did so that their political beliefs are not tampered.

Polls and surveys are a very powerful and useful resource in a wide variety of domains. Therefore, it seems necessary to come up with certain mechanisms, designs, or techniques that incentivize respondents to answer as truthfully as possible. Psychologists and Behavioural Economists have made significant progress in this area developing techniques such as framing effects and randomized response. Stanley Warner (1965) acknowledges that some respondents may choose (for many reasons) to either not respond certain questions or to deliberately give an incorrect response to a certain question for which the participant does not want the survey conductor to know his answer. In his investigation, he argues that this evasive answer bias can be eliminated by using randomized response (Warner, 1965). Bruine de Bruin (2011) concludes that survey responses can be influenced by the wording choice and the presentation order but argues that, although framing has helped survey designers understand how respondents interpret survey questions, decision-making survey results are still not fully reliable (Bruine de Bruin, 2011). It seems that these tools are still far from being the final solution to the problem (Preisendörfer & Wolter, 2014).

There have been, however, newer techniques to incentivize truthful responses for questions for which the veracity cannot be proven by the surveyor. Perhaps the most promising one is the Bayesian truth-serum (BTS) developed by Drazen Prelec (2004). The BTS is a scoring method which rewards participants if they respond truthfully. The method works by asking respondents to make a prediction of how common each answer will be after they have chosen their own answer (Prelec, 2004). After this, respondents are given a score based on how accurate their predictions were and on how common their own response was (Ibid.). This way, the method "rewards responses that are 'surprisingly common' and penalizes those that are 'surprisingly uncommon'" (Weaver & Prelec, 2013, p.289). The BTS method assumes that respondents are Bayesian utility maximisers and the Nash equilibrium is achieved when all participants respond truthfully, and so, maximise their own score (Prelec, 2004). Weaver & Prelec (2013) showed that giving financial incentives according to the participant's BTS score did, in fact, induce truth-telling. Furthermore, they showed that the BTS is more effective than other truth-inducing methods such as the “solemn oath” proposed by Jacquemet et al. (2013) (Weaver & Prelec, 2013). Under the “solemn oath” method, participants are asked to swear on their honour that they will respond truthfully before answering the survey (Jacquemet, Joule, Luchini, & Shogren, 2013). In another research, John, Loewenstein, and Prelec (2012) conducted a survey of academic psychologists in the U.S. and found that BTS truth-telling incentives did, in fact, incentivize participants to confess having done certain questionable research practices, a socially undesired scientific misconduct.

There is one problem, however, under the BTS method for the assumption of the Bayesian Nash equilibrium to hold: “the sample of respondents must be sufficiently large” (Prelec, 2004, p.463). Cvitanić, Prelec, Riley, and Tereick (2019) address this problem in their more recent truth-telling mechanism: Choice-Matching. Similar to BTS, their method attempts to elicit honest responses to multiple-choice questions (MCQ) for which the survey conductor has no way to verify the truthfulness of said responses (Cvitanić et al., 2019). The method consists of adding an *auxiliary question* that asks participants to predict what share of other participants answered each of the possible options of the MCQ (Ibid.). Participants are then given a score based on a weighted average of the accuracy of this prediction and the average score of other participants who answered the same choice from the MCQ (Ibid.). Furthermore, it is not necessary that this additional question consists of making a prediction (Ibid.). This makes Choice-Matching a powerful tool when the survey conductor thinks that having participants make a prediction might influence their own responses. Finally, Cvitanić et al. (2019) argue that their truth-inducing method is easier to explain to respondents than BTS and that their method works with small samples.

There is one particular field in which using real choices might either not be ethical or just not possible, and that is Health Economics. In this field, researchers cannot simply, for example, ask participants to trade life-years for health improvements. However, it is extremely relevant to be able to measure people’s preferences in these kinds of scenarios. The current coronavirus COVID-19 pandemic is an example of how important it is to measure as faithfully as possible people's preferences regarding Health Economics scenarios. During the pandemic, for example, one could ask people if they are willing to make a trade-off between their health and their economy. Simply put, people could be asked if they are willing to go into quarantine for a month, thus protecting their health, but sacrificing part of their income as a result of this quarantine. However, pandemics are a rare event and one cannot produce a pandemic just to measure people's preferences. Therefore, when economists want to measure people's preferences under certain health scenarios, it is likely that they face a *hypothetical bias.* The hypothetical bias is defined as "the difference between hypothetical (stated) and actual (revealed) values" (Özdemir et al., 2009, p.894). Since there is no way of revealing the actual life-years, or days in quarantine traded by participants from the previous example, the hypothetical bias might be larger in health economics than in other fields. As a result, it becomes an interesting field to test truth-telling mechanisms.

In their research, Baillon, Bleichrodt, and Granic (2020) test the Bayesian truth-serum on subjective well-being and health-related questions. In their investigation, the authors compare the effects of the BTS incentives when using a default bias against when there is no default bias in the survey. Their findings suggest that there is no hypothetical bias for health-related questions whether including or not defaults; however, for subjective well-being questions, they conclude that there is a hypothetical bias when including the default bias (Baillon, Bleichrodt, & Granic, 2020). The motivation for this research is, therefore, to test the robustness of these findings by using another method: Choice-Matching.

Chapter Two: Methodology

This chapter consists of multiple sections. The first section explains in more detail the methodology under Cvitanić et al. (2019) choice-matching. The second section explains how the data was collected. The third section describes the design of the survey used to test truth-telling incentives. The final section shows a descriptive summary of the sample.

2.1 Choice Matching

As mentioned briefly in the introduction, choice-matching is a truth-telling mechanism developed by Cvitanić, Prelec, Riley, and Tereick (2019). The method is very similar to Prelec's (2004) Bayesian Truth Serum (BTS). Under the BTS method, respondents are asked to estimate the share of participants who would choose the same answer as they did. To incentivize truth-telling, participants receive a score based on their predictions and get paid according to this score. Participants are told that the methodology for this scoring method makes sure that truthful answers increase the score and thus their payment. Survey designers using BTS try to persuade respondents to answer honestly by the use of "intimidation" (Cvitanić et al., 2019), telling them that they should trust the scoring methodology and the researcher who invented BTS. Specifically, they mention that the BTS was developed by a professor from the MIT, a highly prestigious university, and published on the academic journal *Science*, a highly prestigious journal (Weaver & Prelec, 2013).

Cvitanić et al. (2019) argue that respondents might find BTS' instructions and intuition very complex and even not credible. This is why they developed choice-matching. Under this method, respondents are asked to participate in an *auxiliary game* where they will estimate the share of participants who would choose each of the possible options of the multiple-choice questions (MCQ). Participants are then matched to other respondents who chose the same option from the MCQ. The intuition behind this matching is that people who believe or think the same as one will be more likely to respond in the same manner, making it in the respondents' best interest to respond truthfully so that they are matched with people who think similar to themselves. Without the use of intimidation, participants are told that they will receive a score based on how well they perform on this game and how well participants who they are matched to performed on the game. By using a proper scoring rule, the score is the weighted average between their prediction score and the prediction score of matched participants. It is in the respondents' best interest to respond truthfully since, by doing so, his score will depend on another respondent's prediction which should be very similar to the prediction he or she made (Cvitanić et al. 2019).

Since there is neither intimidation nor the explicit mention in the survey instructions of responding "truthfully", as opposed to the BTS method, choice-matching results in an easier method to be explained to respondents and, furthermore, understood by them. Appendix A shows an example for the instructions under choice-matching.

2.1.1 Example

Imagine that you want to know people's preferences for the upcoming U.S. election, so you design a survey where you ask participants to respond whether they are voting for (1) Donald Trump, (2) Joe Biden, or (3) do not know. Their true answer is called their "type" and is represented by a random vector T^r where $T_i^r = 1$ if i is respondent $r's$ honest answer to the MCQ and $T_i^r = 0$ otherwise. As mentioned in the Literature Review, people might not respond honestly, maybe because they do not want their political beliefs to be tampered with, or they might feel embarrassed of the candidate they support, or they may even feel that they will be judged by the survey conductor for their political preferences. This could evidently damage the quality of the survey's results. This is where the *auxiliary game* and monetary incentives come into play, by asking respondents to also make a prediction of the share of participants who they estimate will choose each answer. Take $x^r = (x_1, x_2, x_3)$ as the answer on the first question of respondent r, where $x_k^r = 1$ if $r's$

answer is k and $x_k^r = 0$ otherwise. For example, if they answered “Joe Biden”, then $x^r = (0,1,0)$. From the additional *auxiliary game*, let $y^r = (y_1^r, y_2^r, y_3^r)$ be the vector of predictions, where each y_k^r represents the respondent’s r prediction of other participants choosing k to the multiple-choice question (MCQ). The respondent will then receive a score based on his own prediction vector y^r and the average score of other respondents who are matched with respondent r because they chose the same answer x^r on the MCQ. Therefore, respondent’s r final score is a weighted average $\lambda S + (1-\lambda)\bar{S}^{-r}$ where S is his own score and $\bar{S}^{-r}$ is the average score of matched respondents excluding respondent r (Cvitanić et al. 2019).

In general, there are N respondents and M possible choices to a MCQ, having $A = \{1, \dots, M\}$ be the set of possible answers to the MCQ. Then, $X^r = (X_1^r, \dots, X_M^r)$ is respondent’s r choice vector and $Y^r = (Y_1^r, \dots, Y_M^r)$ is the prediction vector. The score would then be obtained similarly to the above example.

2.1.2 Assumptions

Cvitanić et al. (2019) list the following four assumptions for choice-matching to be honesty-inducing (Cvitanić et al. 2019, p.182):

- **ASSUMPTION 1** (Common Prior): *There exists a common prior on the distribution of $T^1, \dots, T^N$.*
- **ASSUMPTION 2** (Non-Degeneracy): *For any respondent r and any realization t^r: $\Pr(\mathcal{E}^r | T = t^r) > 0$.*
  Where $\mathcal{E}^r$ is the type-matching trigger of respondent r, that is, “the event such that each type $i \in A$ is represented at least once among respondents other than r” (Cvitanić et al. 2019, p.181).
  This assumption suggests that “each respondent considers it possible that all distinct types are represented in the rest of the sample” (Cvitanić et al. 2019, p.182).

- **ASSUMPTION 3** (Stochastic Relevance): *For any two respondents* r, s *and any answer options* $k, i \in A$*:*

$$p^{r,k} \neq p^{s,i} \; if \; k \neq i$$

Where $p^{r,k} = E[\bar{T}^{-r} | T_k^r = 1, \mathcal{E}^r]$.

- **ASSUMPTION 4** (Impersonal Updating): *For any two respondents* r, s *and any answer option* $k \in A$*:*

$$p^{r,k} = p^{s,k}$$

2.2 Data

The data used for this thesis was collected by doing an online experiment, namely, an online survey. The survey was constructed using Qualtrics XM. The distribution of the survey was made through an anonymous link, shared among students at Erasmus University Rotterdam and posted on social media. The survey had 226 participants uniformly distributed between the control group and the treatment (incentives) group. However, only a total of 174 participants completed the survey, 107 from the control group and 67 from the treatment group. The median of time spent in the survey was 10 minutes, which could be considered as an excessive amount of time for taking a survey for some participants, causing them to drop out. The fact that more participants dropped out from the treatment group than from the control group could be due to the following three reasons: (1) the instructions for the treatment group were longer and could have been perceived as more complex or difficult to understand, (2) some participants might be intimidated by having to "compete" in a game, and (3) participants in the treatment group were asked for their email to receive a €10 prize; although providing the email was optional, some participants might have misunderstood and thought that providing their email was mandatory and thus decided to drop out.

2.3 Survey

The survey included the following sections: Introduction and instructions, subjective well-being questions, health-related hypothetical questions, and subjective questions related to respondents' behaviours and beliefs during the coronavirus (COVID-19) outbreak. Both the subjective well-being and health-related sections are similarly included in Baillon, Bleichrodt, and Granic (2020)'s research, hence these sections were used to test the robustness of their findings. Additionally, health questions are particularly relevant for economists, since these types of questions may entail hypothetical biases. The coronavirus section was included since the pandemic was just beginning in Europe at the time the survey was sent, making it interesting to elicit people's behaviours and beliefs during such a rare and impactful event. Respondents' demographics were asked at the end of the survey. The complete transcript of the survey can be seen in **Appendix A**.

Participants were randomly assigned to either the control group or the treatment group. Instructions in the control group only politely asked participants to respond truthfully, as their answers would be important for the research. Instructions for the treatment group explained that participants would receive a score based on their answers and other respondents' answers, and that the three participants with the highest scores would receive a €10 prize. The instructions were as follows:

Control group: *Your answers are important for my research, please answer truthfully.*

Treatment group: *You will receive a Score based on:*

(1) The accuracy of your predictions on the game and
(2) The accuracy of the predictions of other respondents who chose the same options as you did for the multiple-choice questions.

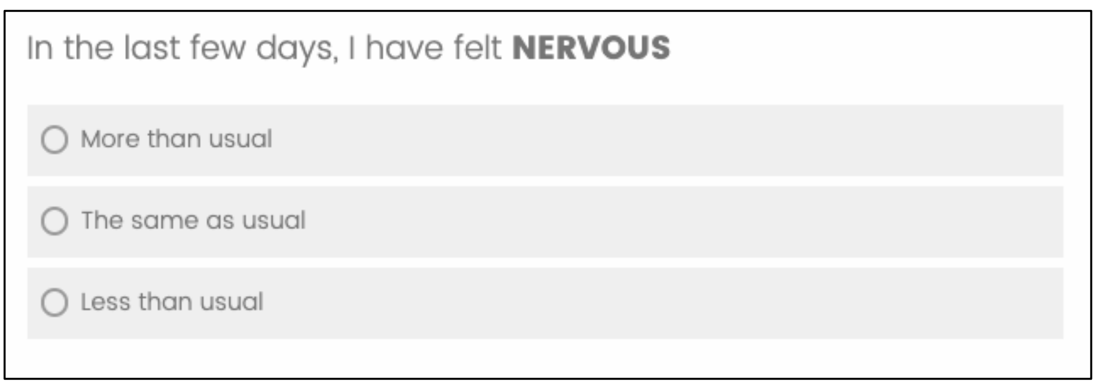
The three participants with the highest Score will be contacted via email and will receive a €10 prize.

The full introduction and instructions can be seen in **Appendix A**.

2.3.1 Subjective well-being

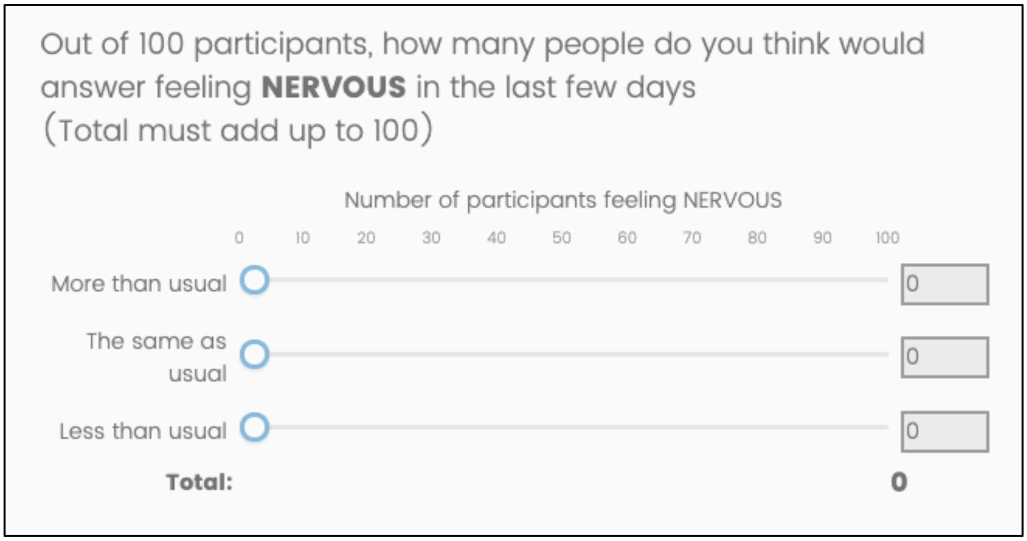
After the instructions, participants were presented with the **subjective well-being** section. This section included five questions asking respondents to answer how they had felt during the last few days. The survey was distributed during the last weeks of April 2020, one month after the quarantine for the coronavirus outbreak had been implemented and was still going on. Specifically, participants were asked if they had felt nervous, stressed, depressed, fatigued, and had trouble concentrating less than usual, the same as usual, or more than usual "during the last few days". **Figure 1** shows an example from this section.

Figure 1: Subjective well-being example



After completing the subjective well-being section, participants were asked to participate in the *auxiliary question* (or *game*, as it was described to the participants) for each of the subjective feelings. They were asked to make a prediction of how many participants, out of 100, would choose each of the possible options for the multiple-choice questions. This was done by either using a slider or introducing their estimation on the grey boxes. The total was forced to sum up to 100. **Figure 2** shows an example from this section.

Figure 2: Subjective well-being auxiliary question



2.3.2 Health

The second part of the survey was the **Health** section. This section included one health scenario and three *standard gamble* (SG) questions related to the same health scenario. A standard gamble question asks participants to choose between living in a certain health condition for the rest of their lives or taking a risky medical procedure with two possible outcomes: probability p of being successful and thus living for the rest of their lives in full health, and probability $(1-p)$ of not being successful and resulting in their death. These types of questions are widely used in health research to measure people's life quality (Dolan, Gudex, Kind, & Williams, 1996). The health states description was taken from the EuroQol EQ-5D-5L (nd), a commonly used reference for health economics. **Figure 3** shows the health conditions used for this section.

Figure 3: Health Conditions

Consider the following health states:

Health State A:

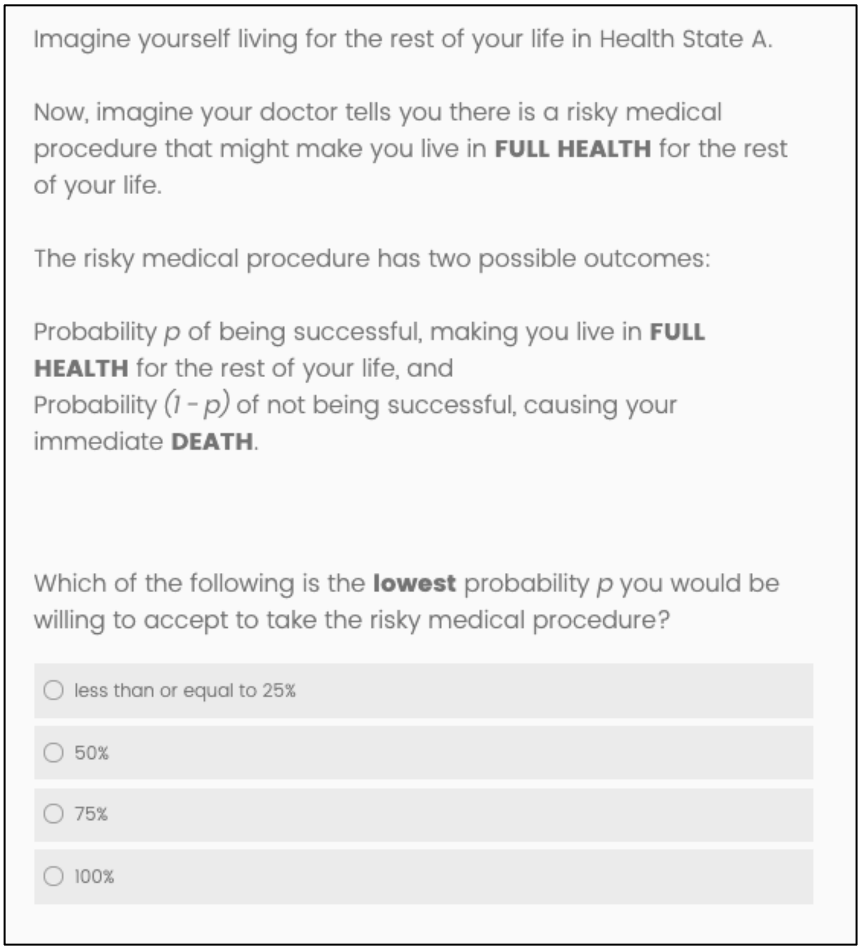
- Unable to walk
- Moderate trouble with self-care activities (e.g. washing or dressing)
- Moderate problems doing usual activities (work, study, family, or leisure)
- Severe pain or discomfort
- Severely anxious or depressed

Full Health:

- No problems to walk
- No trouble with self-care activities
- No problems doing usual activities
- No pain or discomfort
- Not anxious or depressed

Participants were asked which is the lowest probability p they would be willing to accept to take the risky medical procedure. Four different probabilities were presented as possible options: *less than or equal to 25%, 50%, 75%, or 100%*. **Figure 4** shows the instructions for this section.

Figure 4: Health section question



Imagine yourself living for the rest of your life in Health State A.

Now, imagine your doctor tells you there is a risky medical procedure that might make you live in **FULL HEALTH** for the rest of your life.

The risky medical procedure has two possible outcomes:

Probability *p* of being successful, making you live in **FULL HEALTH** for the rest of your life, and
Probability *(1 - p)* of not being successful, causing your immediate **DEATH**.

Which of the following is the **lowest** probability *p* you would be willing to accept to take the risky medical procedure?

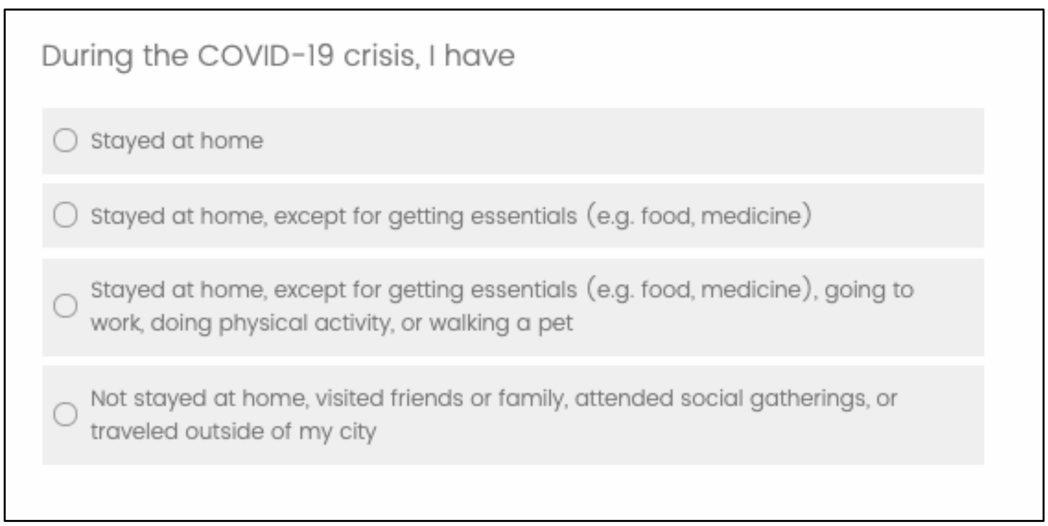
- less than or equal to 25%
- 50%
- 75%
- 100%

Participants were then presented with the *auxiliary question* and asked to estimate the number of participants who would choose each of the possible probabilities. After answering the *auxiliary question*, participants were presented with another SG question considering the same health conditions. The probabilities of this new SG question were linked to their answer of the first SG question. For example, if a participant chose *less than or equal to 25%*, the second SG question would present the following probabilities: *6%, 12%, 18%, and 25%.* A third and final SG question was included for which the probabilities were linked to the second SG question. For example, if a participant chose *12%* on the second SG question, then the third SG question presented the following probabilities: *8%, 10%, and 12%.* This breakdown of probabilities was intended to find the lowest probability participants were willing to accept in order to undertake the risky medical procedure. There were no *auxiliary questions* included for these additional SGs. **Appendix A** shows an example for the chaining of probabilities.

2.3.3 Behaviour and beliefs during the coronavirus outbreak

The final section of the survey asked four questions related to participants' behaviours and beliefs during the coronavirus (COVID-19) outbreak. Specifically, participants were asked about their beliefs on their governments' reaction to the outbreak, whether they had stayed at home or not, whether they had washed their hands more frequently than before the outbreak, and whether they had presented any symptoms. Respondents were asked to respond the *auxiliary question* for each of the four questions in this section. **Figure 5** shows an example question from this section. **Appendix A** shows the full list of questions for this section.

Figure 5: Example from behaviour and beliefs section


During the COVID-19 crisis, I have

- ○ Stayed at home
- ○ Stayed at home, except for getting essentials (e.g. food, medicine)
- ○ Stayed at home, except for getting essentials (e.g. food, medicine), going to work, doing physical activity, or walking a pet
- ○ Not stayed at home, visited friends or family, attended social gatherings, or traveled outside of my city

2.4 Summary Statistics

The total sample size obtained was made of 174 participants, 107 assigned to the control group and 67 assigned to the treatment (incentives) group. The average age for the whole sample was between 25 to 34 years old. Only around 30% of the participants were students. **Table 1** shows the descriptive statistics of the sample.

Table 1: Descriptive statistics of the sample

Concept		Control	Treatment	Total
Number of observations		107	67	174
Average age		25 - 34	25 - 34	25 - 34
Gender				
	Male	*49.5%*	*53.7%*	*51.1%*
	Female	*49.5%*	*44.8%*	*47.7%*
	Non-binary	*0.0%*	*1.5%*	*0.6%*
	Not disclosed	*0.9%*	*0.0%*	*0.6%*
Job status				
	Student	*31.8%*	*22.4%*	*28.2%*
	Employed	*59.8%*	*70.1%*	*63.8%*
	Unemployed	*8.4%*	*7.5%*	*8.0%*
Country of birth				
	Asia / Pacific	*3.7%*	*0.0%*	*2.3%*
	Europe	*11.2%*	*13.4%*	*12.1%*
	Latin America	*71.0%*	*73.1%*	*71.8%*
	Middle East	*0.9%*	*0.0%*	*0.6%*
	North America	*13.1%*	*13.4%*	*13.2%*
Minutes spent in survey				
	Median	9.4	10.6	10.1
	75th percentile	12.7	16.0	14.5

Chapter Three: Results

3.1 Variable Description

Table B.1 included in **Appendix B** shows the full list of variables that were collected and used for the analysis of the results discussed in the following section.

3.2 Analysis of the Results

3.2.1 Effects of Incentives

The first step was to test whether incentives had an effect on the prediction score of the participants. Having a better score, therefore being more accurate in their predictions, can be a result of people concentrating and thinking harder. It can also be an indication that people were overall responding truthfully, according to the choice-matching methodology. It is therefore interesting to test whether incentives had an effect or not on the participants' score. The effect of incentives can be studied by comparing the results from the control group and the incentives group. First, the effect of incentives on the participants' score was studied using a linear regression. The individual score, referred to as the 'prediction score' by Cvitanić et al. (2019), was measured as the squared difference between the participant's prediction and the actual value. This implies that a lower score means that the prediction was more accurate and, thus, better. The final score was obtained as the weighted average of the prediction score and the average prediction score of matched participants. The regression showed that being in the incentives group was associated with making better predictions (lower scores), although this effect was not statistically significant at a 10% level of significance. This result suggests that participants in the treatment group might be incentivized to think harder about their predictions, making them more accurate. Results of the regression can be seen in **Table C.1** in **Appendix C**. Additionally, a student t-test shows that there is evidence (at a 10% significance level) to reject the null hypothesis that the mean scores between the two groups are equal. This suggests that the mean score between the control and the treatment groups was indeed different. **Table 2** shows the student t-test results.

Table 2: Two-sample t test for score

Group	Observations	Mean	Std. Err.	Std. Dev.
Control	107	2.330	0.054	0.557
Incentives	67	2.176	0.062	0.510
combined	174	2.270	0.041	0.543
difference		0.154	0.084	

$H_0: \mu_{control} - \mu_{incentives} = 0$
t-value = 1.8311
p-value = 0.069

Subjective well-being section

The subjective feelings included in the subjective well-being section were negative emotions, namely: feeling nervous, stressed, depressed, tired, and having trouble concentrating. During a pandemic, these kinds of "negative" feelings may arise and become more present among individuals. It was therefore interesting to investigate the degree to which the pandemic had accentuated these kinds of emotions. Normally, however, people tend to overreport positive emotions (Baillon, Bleichrodt, & Granic, 2020). Therefore, it would be expected to see people in the control group choosing the "less than usual" and "same as usual" options more often than people in the incentives group. However, this was not the case as shown later in section 3.2.2, **Figure 8**.

To test whether incentives had an effect, ordered probit regressions were run for each of the subjective emotions. The results of these regressions show a positive effect of incentives for all emotions except feeling nervous. This means that participants in the incentives group were more likely to report "more than usual" for all subjective feelings except nervous. However, this effect was not statistically significant at a 10% level of significance for any of the regressions and, therefore, causality cannot be inferred for this effect. **Table C.2** in **Appendix C** shows the results of the ordered probit regressions for the subjective well-being section.

Finally, Mann-Whitney U tests were run for each subjective feeling to test whether there was a difference in the average response between the control group and the treatment group. Results suggest that there was no evidence to reject the null hypothesis that the mean for the subjective feeling differed between the control and treatment groups. **Table D.1** in **Appendix D** shows the results for the Mann-Whitney U tests for this section. Therefore, it seems that incentives had no effects in the subjective well-being section.

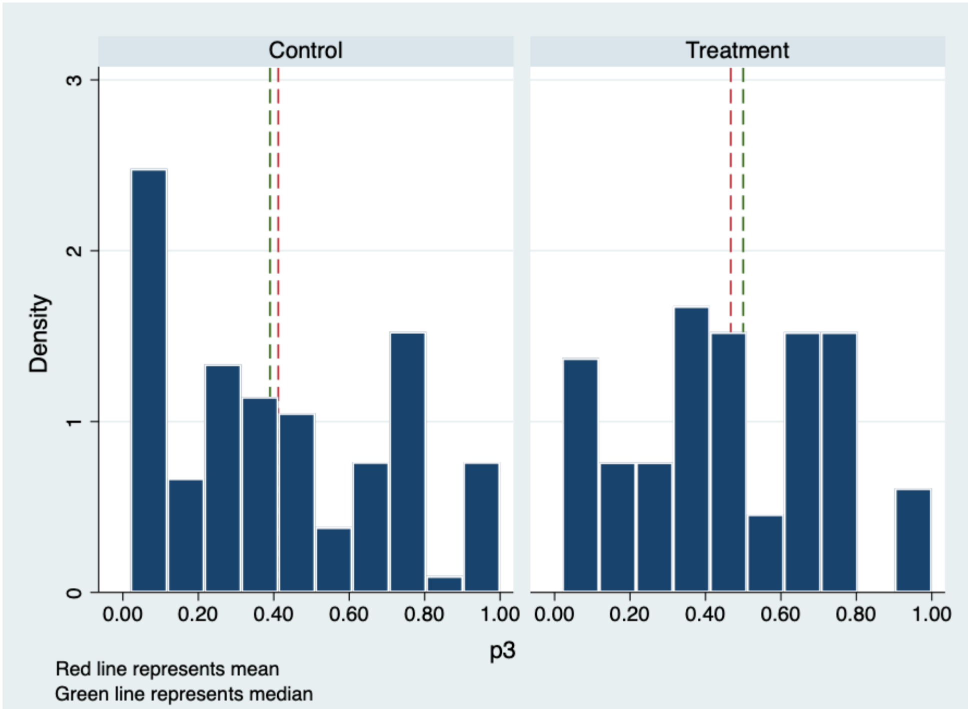
Health section

The Health section asked participants to choose the lowest probability they were willing to accept in order to take a risky medical procedure while living in Health State A (as described in **Figure 3** from Chapter 2). A first probability p_1 was chosen by participants. Depending on their choice for p_1, linked options appeared for p_2 and finally for p_3, simulating this way a breakdown of probabilities in order to find out the lowest probability participants were willing to accept.

Linear regressions were run for the reported probabilities $p_1, p_2, and\ p_3$ to test whether incentives had an effect on the lowest probability participants would be willing to accept in order to undertake the medical procedure. Incentives were positively associated with the probabilities p_i, but this effect was not statistically significant at a 10% level of significance. This suggests association but not a causal effect between introducing incentives and choosing the probabilities. Results for the linear regressions can be seen in **Table C.3** in **Appendix C**.

Since p_3 represents the lowest probability participants were willing to accept, testing for differences between the control group and the incentives group in their choice for p_3 becomes of interest. The average probability p_3 was *41.2%* and *46.7%* for the control and incentives group respectively. **Figure 6** shows the histogram of p_3 by group; the mean is represented by a red line and the median by a green line.

Figure 6: Histogram of p3 by group



A Mann-Whitney U test and a t-test were run to see whether the average probability was statistically different between groups. Results show that there is no evidence to reject the null hypothesis that the mean between the two groups is different, suggesting that incentives did not have an effect. **Table 3** shows the Mann-Whitney U test and t-test results for p_3. Additional Mann-Whitney U tests were run for p_1 and p_2, but results showed no statistical difference between the two groups either. **Table D.2** in **Appendix D** shows the results for the Mann-Whitney U tests for the health section. Overall it seems that incentives had no effects on health-related questions.

Table 3: Two-sample Mann-Whitney U test and t-test for p3

Mann-Whitney U test p3

Group	Obs	Rank Sum	Expected
Control	107	8,945.0	9,362.5
Incentives	67	6,280.0	5,862.5
combined	174	15,225.0	15,225.0
H_0: p3(incentives==0) = p3(incentives==1)			
z	-1.294		
p-value	0.1956		

t-test p3

Group	Obs	Mean	Std. Err.	Std. Dev.
Control	107	0.412	0.029	0.296
Incentives	67	0.467	0.033	0.266
combined	174	0.433	0.022	0.285
difference	-	0.055	0.044	
H_0: Mean difference = 0				
t-value	-1.240			
p-value	0.2165			

Behaviour and beliefs during the coronavirus outbreak section

The coronavirus section involved questions about people's beliefs and behaviours during the pandemic. Washing your hands and staying at home might be seen as socially desirable behaviours during the pandemic. Therefore, it would be expected that people choose the most socially desirable options: handwashing more frequently than usual and staying at home. It would also be expected that introducing incentives would incentivize people to choose more "relaxed" options, so to speak. Presenting COVID-19 symptoms might be seen by some people as not desirable or even embarrassing. Therefore, it would be expected to see most people stating that they have not presented any symptoms. It would then be expected that, by introducing incentives, people would be more likely to report that they had indeed presented symptoms. An ordered probit regression was run for each of the questions in this section. The results of these regressions can be seen in **Table C.4** in **Appendix C**.

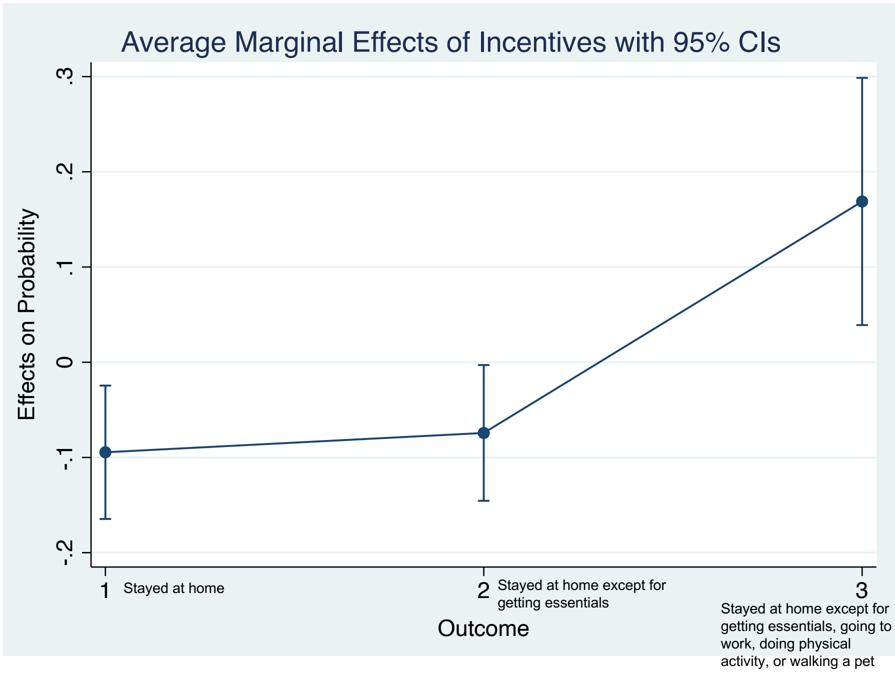
Incentives had no significant effects for all questions except for the question about staying at home. The use of incentives had a positive and significant coefficient (at a 5% level of significance). Average marginal effects were then obtained for this particular question to see the magnitude of the effect of incentives. Results suggest that using incentives increase the probability of choosing option 3 (“Stayed at home, except for getting essentials, going to work, doing physical activity, or walking a pet”) by 16.9 percentage points, ceteris paribus. This effect is statistically significant at a 5% level of significance. This means that introducing incentives actually incentivized people to confess doing a more “relaxed” quarantine, which can be seen by some people as a not socially desirable behaviour. This result is in line with John, Loewenstein, and Prelec’s (2012) findings discussed in the Literature Review, that incentives can motivate people to report socially undesired behaviour. **Table 4** and **Figure 7** show the marginal effects of incentives on answering the question about staying home during the pandemic.

Table 4: Average marginal effects of Incentives for question "stay home"

Group	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf. Interval]	
control	(base outcome)					
incentives						
outcome						
1	-0.095	0.036	-2.65	0.008	-0.165	-0.025
2	-0.074	0.036	-2.04	0.041	-0.146	-0.003
3	0.169	0.066	2.55	0.011	0.039	0.299

Note: dy/dx for factor levels is the discrete change from the base level

Figure 7: Average marginal effects of Incentives for question "stay home"



Finally, Mann-Whitney U tests were run to see if there was a difference on average between groups for each of the questions in this section. Again, only the “staying home” question showed evidence for a statistically significant difference. Formally, there was evidence to reject the null hypothesis that the average response for the “staying home” question in the control group was different to the average response in the incentives group. An additional student t-test was run for robustness. This result holds for both the Mann-Whitney U and the t-test as shown in **Table 5**. The results for the rest of the tests performed in this section can be seen in **Table D.3** in **Appendix D**.

Table 5: Two-sample Mann-Whitney U test and t-test for question "stay home"

Mann-Whitney U test stay_home

Group	Obs	Rank Sum	Expected
Control	107	8,577.5	9,362.5
Incentives	67	6,647.5	5,862.5
combined	174	15,225.0	15,225.0
H_0: stay_home(incentives==0) = stay_home(incentives==1)			
z	-2.713		
p-value	0.0067		

ttest stay_home

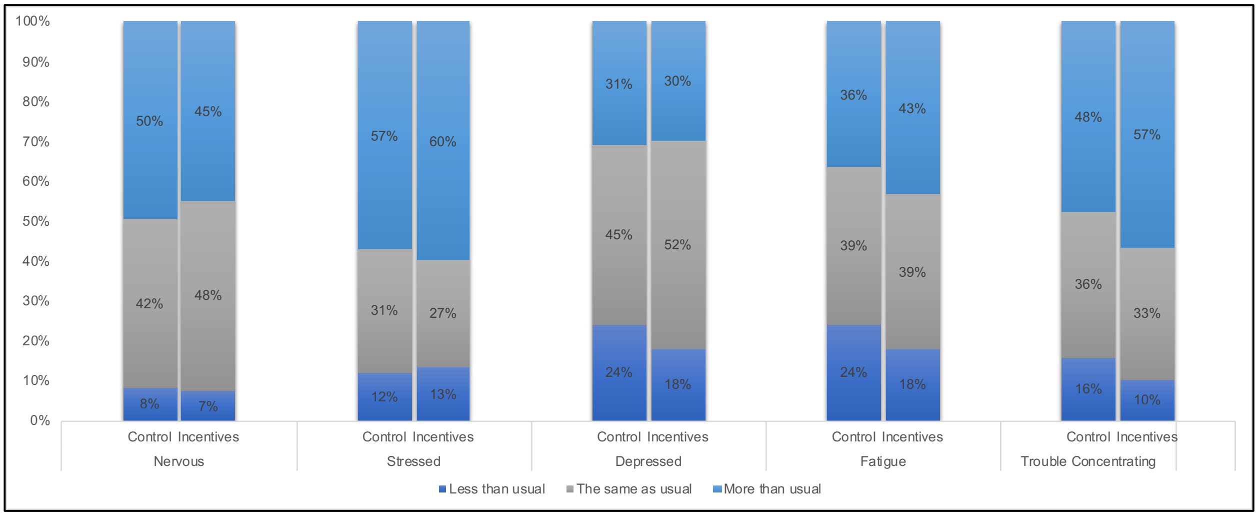
Group	Obs	Mean	Std. Err.	Std. Dev.
Control	107	2.056	0.062	0.642
Incentives	67	2.328	0.078	0.637
combined	174	2.161	0.049	0.652
difference	-	0.272	0.100	
H_0: Mean difference = 0				
t-value	-2.731			
p-value	0.0070			

3.2.2 Additional Findings

Subjective well-being section

Half of the participants reported feeling more nervous than usual during the coronavirus outbreak. Around 60% reported feeling more stressed than usual. Only about 30% of participants reported feeling more depressed than usual during the pandemic. These results hold for both the control group and the treatment group. Regarding feeling fatigue, 36% of participants in the control group and 43% of participants in the treatment group reported feeling more tired than usual. Half of the participants in the control group and almost 60% of participants in the treatment group reported having more trouble concentrating during the pandemic. **Figure 8** shows the distribution of participants for the subjective well-being section. Feeling more negative emotions during a pandemic is expected and, therefore, data quality appears to be good for both groups.

Figure 8: Distribution of participants in the subjective well-being section



Correlation matrixes were obtained for both the control and the incentives groups. For the control group, being male was positively correlated with feeling more nervous, stressed, and having trouble concentrating; and negatively correlated with feeling depressed and tired. For the incentives group, however, being male was negatively correlated with feeling nervous, stressed, depressed, and having trouble concentrating; and positively correlated with feeling tired. The opposite held for women. For the control group, statistically significant correlations (at a 5% level of significance) were a negative correlation between being married / co-habiting and having trouble concentrating; and a negative correlation between other gender and feeling nervous or stressed. For the incentives group, statistically significant correlations (at a 5% level of significance) were the following: age, being married / co-habiting, and being employed were negatively correlated with having trouble concentrating; male and other gender were negatively correlated with feeling nervous; female was positively correlated with feeling nervous; being from the EMEA region and being a student were positively correlated with having trouble concentrating; and being from North America was positively correlated with feeling depressed. **Table 6** shows the correlation matrixes between subjective feelings and demographic characteristics first for the control group and then for the incentives group.

Table 6: Correlation matrix between subjective feelings and demographic characteristics

Control Group

	nervous	stressed	depressed	fatigue	concentrating
nervous	1				
stressed	0.5263*	1			
depressed	0.1207	0.2140*	1		
fatigue	0.3156*	0.3843*	0.1502	1	
concentrating	0.18	0.1414	0.1689	0.031	1
age_cat	0.0349	-0.0287	0.0244	-0.0713	-0.039
male	0.0644	0.006	-0.0624	-0.0592	0.0296
female	-0.0232	0.0327	0.064	0.0865	-0.0215
other_gender	-0.2140*	-0.2009*	-0.0086	-0.1415	-0.0422
Asia	0.0273	0.0848	0.0492	-0.0951	0.1165
EMEA	-0.0154	0.1294	-0.0329	0.1271	-0.1225
LatinAmerica	-0.1046	-0.091	0.0843	0.0205	0.024
NorthAmerica	0.1403	-0.0507	-0.1093	-0.0972	0.0209
education	0.0526	0.0525	-0.1178	-0.0317	-0.0439
married	0.0642	0.0167	-0.0437	0.1003	-0.2637*
student	-0.0621	-0.0072	-0.0875	-0.0034	0.0054
employed	0.1096	0.0079	0.0467	0.0056	0.0955
unemployed	-0.0894	-0.0018	0.0642	-0.0041	-0.1778

Note. * $p<0.05$

Incentives Group

	nervous	stressed	depressed	fatigue	concentrating
nervous	1				
stressed	0.4503*	1			
depressed	0.1776	0.2225	1		
fatigue	-0.1089	0.0879	0.2658*	1	
concentrating	0.0867	0.2963*	0.2690*	0.1531	1
age_cat	0.1444	-0.1439	-0.0305	-0.1351	-0.2889*
male	-0.2627*	-0.0273	-0.1451	0.1159	-0.2061
female	0.3300*	0.0467	0.1066	-0.1465	0.1828
other_gender	-0.2731*	-0.0792	0.1592	0.1241	0.0978
Asia	.	.	.	.	.
EMEA	-0.096	0.1726	-0.0048	-0.0168	0.2482*
LatinAmerica	-0.0154	-0.1719	-0.1905	-0.0652	-0.2325
NorthAmerica	0.1161	0.0509	0.2524*	0.1015	0.0541
education	0.0613	0.1089	-0.0975	-0.0214	0.027
married	0.081	-0.0569	-0.0665	0.0408	-0.3015*
student	-0.0345	0.1523	0.1162	0.1546	0.3737*
employed	0.0244	-0.1245	-0.1252	-0.2171	-0.2771*
unemployed	0.0123	-0.0248	0.0336	0.1329	-0.1103

Note. * $p<0.05$

Health section

For the health section, roughly 90% of the full sample was evenly distributed over the *25%, 50%,* and *75%* options for p as the lowest probability they were willing to choose to undertake the risky medical procedure in the first question of the section (p_1). The majority of the control group chose *25%,* while the majority of the incentives group chose *50%* as the minimum probability p for the risky medical procedure. **Figure 9** shows the distribution of participants for the first question (p_1) of the health section. **Figure 6** in section 3.2.1 showed the histogram of p_3 by group. As mentioned in that section, the mean probabilities for the control group and the incentives group were *41.2%* and *46.7%* respectively. These probabilities suggest that participants preferred avoiding the chance of dying even when living under the conditions of Health State A. Although these probabilities may seem relatively high, previous research has found probabilities for Standard Gamble questions to be greater than *50%* (Van Osch, Wakker, Van Den Hout, & Stiggelbout, 2004; Van Osch & Stiggelbout, 2008). Regarding probabilities p_2 and p_3, the dispersion for the chosen probabilities was smaller in the incentives group than in the control group, as shown in **Figure 10**. This may be an indication that incentives made participants focus more on the breakdown questions and think harder of their responses.

Figure 9: Distribution of participants in the health section, question p_1

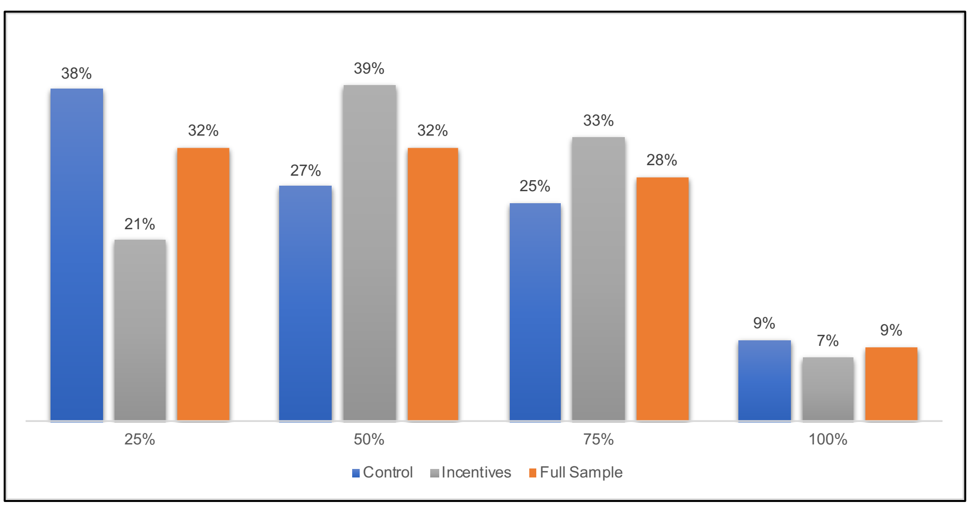
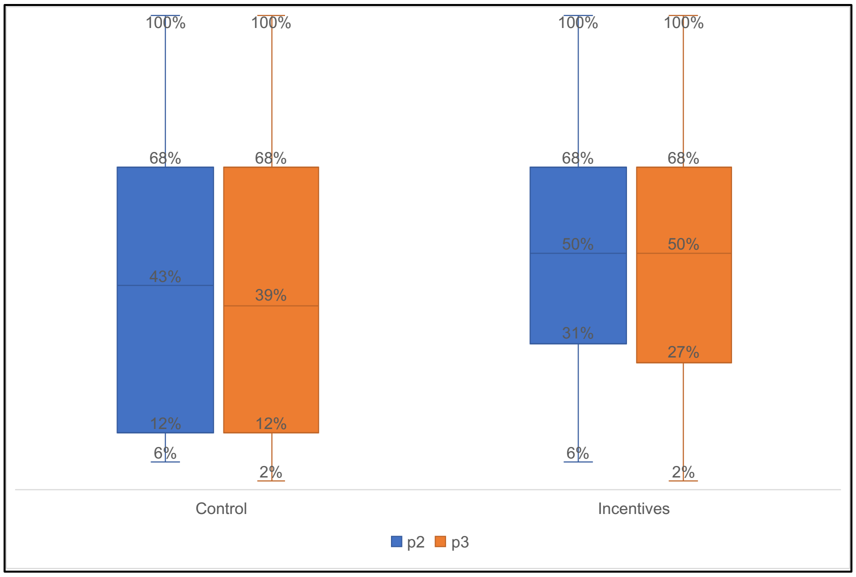


Figure 10: Boxplot of probabilities p2 and p3



Correlation matrixes were obtained for both the control and the incentives groups. For the control group, being a male was associated with choosing higher probabilities for taking the risky medical procedure in the three questions p_1, p_2 and p_3, which would contradict previous research that, on average, men are more risk-seeking than women. On the incentives group, however, being a male was indeed correlated with accepting lower probabilities, therefore accepting more risk, which is in line with previous research. This could suggest that introducing incentives might induce truth-telling since it is seen that, by using them, results for gender are in line with previous thoroughly explored research. However, none of these correlations were statistically significant at a 5% level of significance. For the control group, the only statistically significant correlations (at a 5% level of significance) were the following: being a student was correlated with choosing a higher probability and being unemployed / retired was correlated with choosing a lower probability, therefore accepting higher risk. For the incentives group, there were no statistically significant correlations (at a 5% significance level) between chosen probabilities and demographic characteristics. Although not statistically significant, education was positively correlated with the reported probabilities in the incentives group. Previous research has shown that more educated people are willing to accept more risk;

therefore, this result would not be in line with previous research. **Table 7** shows the correlation matrixes between probabilities and demographic characteristics for both groups.

Table 7: Correlation matrix between probabilities p_i and demographic characteristics

Control Group

	p1	p2	p3
p1	1		
p2	0.9665*	1	
p3	0.9597*	0.9986*	1
age_cat	-0.1648	-0.1483	-0.1515
male	0.0751	0.0249	0.0226
female	-0.0554	-0.001	0.0023
other_gender	-0.1023	-0.1243	-0.1295
Asia	0.0381	0.0354	0.0354
EMEA	0.1218	0.1643	0.1694
LatinAmerica	-0.1491	-0.1521	-0.1571
NorthAmerica	0.0612	0.0255	0.0273
education	-0.0007	0.009	0.0111
married	-0.1016	-0.1207	-0.1339
student	0.182	0.1932*	0.1963*
employed	-0.0492	-0.0461	-0.0497
unemployed	-0.2183*	-0.2426*	-0.2415*

Note. * p<0.05

Incentives Group

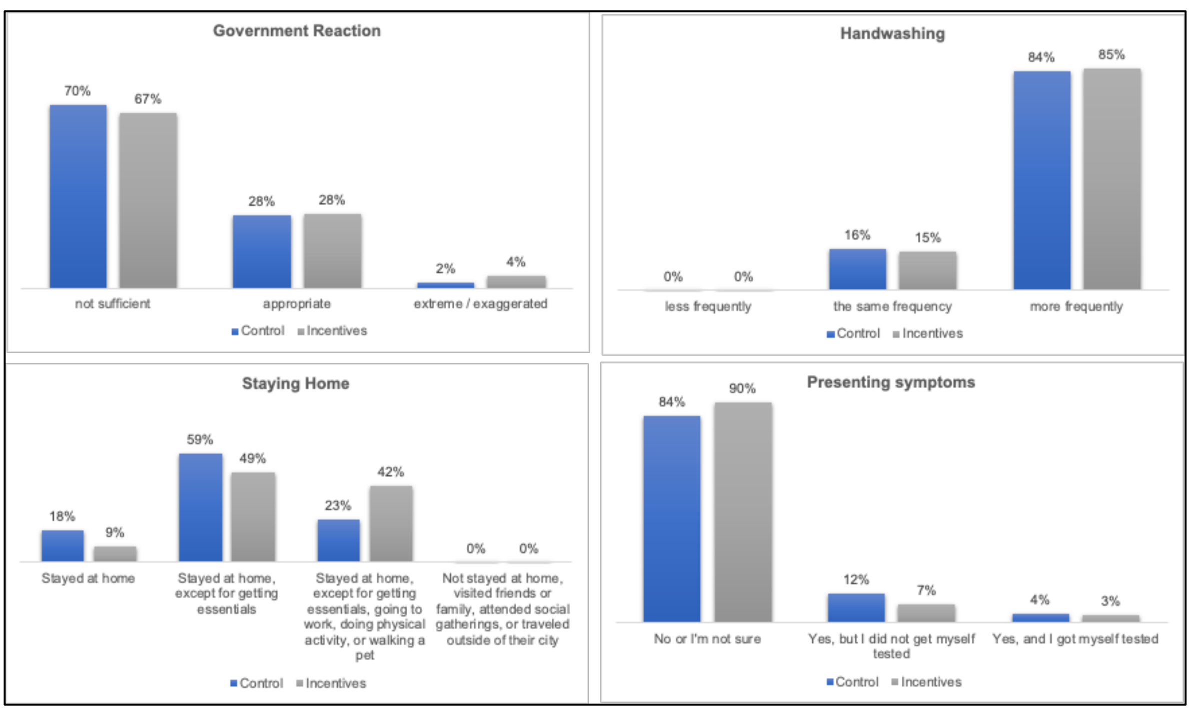
	p1	p2	p3
p1	1		
p2	0.9664*	1	
p3	0.9574*	0.9982*	1
age_cat	-0.0309	-0.0323	-0.0405
male	-0.1943	-0.1837	-0.1798
female	0.204	0.2044	0.2004
other_gender	-0.0378	-0.0826	-0.0825
Asia	.	.	.
EMEA	0.0292	0.0321	0.0406
LatinAmerica	0.1478	0.1224	0.1221
NorthAmerica	-0.2213	-0.1912	-0.1994
education	0.1844	0.1524	0.138
married	0.0607	0.0994	0.0847
student	-0.0012	-0.0078	-0.001
employed	0.0139	0.0213	0.0201
unemployed	-0.0223	-0.0248	-0.0335

Note. * p<0.05

Beliefs and behaviours during the coronavirus outbreak section

Figure 11 shows the results obtained from the last section of the survey, related to beliefs and behaviours during the pandemic. Around 70% of the participants perceived their government's reaction to the pandemic as not sufficient. 85% of participants reported washing their hands more frequently than before the outbreak, while not a single participant reported washing their hands less frequently than before the outbreak. This was expected since, at that moment, a worldwide campaign asking people to thoroughly wash their hands for at least 20 seconds was in place. No participant reported breaking the quarantine by going outside of their homes to visit friends or family or traveling outside their cities. Around 90% of participants reported not having (or not being sure about having) any COVID-19 symptoms. This was expected since large-scale COVID-19 testing was not in place at the moment.

Figure 11: Summary of results from beliefs and behaviours during the coronavirus outbreak



As in the other sections, correlation matrixes were obtained for both the control and the incentives groups. Although not statistically significant (at a 5% level of significance), being a male was negatively correlated to handwashing in both groups. This could mean that women are more responsible than men about taking care of their health. Being a male was correlated with *stay_home*, meaning that men were associated with more flexible quarantining; this correlation was not statistically significant either. This is also in line with women being more responsible and socially minded than men. In the control group, statistically significant correlations (at a 5% level of significance) were the following: Being from the EMEA region and being a student were both correlated with perceiving their government's reaction as extreme / exaggerated, while being from Latin America and being employed were both correlated with perceiving their government's reaction as insufficient; being employed was correlated with not having coronavirus symptoms while being unemployed showed the opposite. Regarding the incentives group, statistically significant correlations (at a 5% level of significance) were the following: other gender and being from the EMEA region were correlated with perceiving their government's reaction as extreme / exaggerated and negatively correlated with handwashing; being either a student or unemployed were correlated with presenting coronavirus symptoms, while being employed was correlated with not presenting coronavirus symptoms. **Table 8** shows the correlation matrixes between the coronavirus questions and demographic characteristics for both groups.

Table 8: Correlation matrix between coronavirus questions and demographic characteristics

Control Group

	gov_react	handwash	stay_home	covid_symp~s
gov_react	1			
stay_home	-0.1318	1		
handwash	0.2058*	-0.1219	1	
covid_symp~s	0.1664	0.0709	0.0856	1
male	0.1542	-0.0808	0.0593	0.062
female	-0.1424	0.0726	-0.0577	-0.0543
other_gender	-0.0612	0.0422	-0.0085	-0.0395
Asia	0.0712	-0.0491	0.0598	0.1241
EMEA	0.3330*	0.0834	0.1465	-0.0327
LatinAmerica	-0.3739*	-0.0521	-0.1697	-0.0391
NorthAmerica	0.1403	0.017	0.0527	0.0145
education	0.1041	-0.0573	0.0651	0.0717
married	0.0371	0.0796	-0.1595	-0.0417
student	0.2865*	0.1319	0.0972	0.0552
employed	-0.2774*	-0.1477	-0.0176	-0.2198*
unemployed	0.0094	0.0396	-0.132	0.2956*

Note. * $p<0.05$

Incentives Group

	gov_react	handwash	stay_home	covid_symp~s
gov_react	1			
stay_home	-0.1671	1		
handwash	0.0744	0.085	1	
covid_symp~s	0.1653	-0.1654	-0.11	1
male	0.0299	-0.1367	0.0085	-0.0596
female	-0.1158	0.2087	0.0071	0.0694
other_gender	0.3521*	-0.2939*	-0.0639	-0.0394
Asia	.	.	.	.
EMEA	0.3572*	-0.0807	-0.1354	0.0825
LatinAmerica	-0.1352	0.1241	-0.0048	0.0335
NorthAmerica	-0.1815	-0.0807	0.1416	-0.1261
education	0.1724	0.1485	0.0123	-0.1661
married	0.0308	-0.0014	0.0739	-0.0534
student	-0.0376	-0.0765	-0.1091	0.2548*
employed	0.0265	0.0929	0.1325	-0.4131*
unemployed	0.0134	-0.0404	-0.0577	0.3152*

Note. * $p<0.05$

3.3 Discussion

Recalling the research question:

Do truth-telling incentives have an effect on questions about subjective well-being, health, and beliefs & behaviour?

The answer seems to be no; at least not for subjective well-being and health-related questions. Results suggest that data quality was good overall for both incentivized and hypothetical questions. This result is in line with Baillon, Bleichrodt, and Granic (2020)'s findings, who used the BTS methodology on similar questions. The authors found no effect of incentives on health-related questions regardless of whether default bias was included or not. For subjective well-being questions, however, they did find an effect of incentives but only when the default bias was used. Since the questions in this investigation did not include a default bias or another type of cognitive bias, the results suggest that even using a different methodology, in this case choice-matching, truth-telling incentives have no effect on subjective well-being and health-related questions. Therefore, it seems that there is actually no hypothetical bias in these types of questions, and it is then the responsibility of the survey designer to ensure data quality and avoid other kinds of biases (Baillon, Bleichrodt, & Granic, 2020).

However, choice-matching incentives did have a statistically significant effect on one particular question. This question was in the last section of the survey and it was related to behaviour during the coronavirus pandemic. The question was specifically related to how strict people were regarding being in quarantine, by staying home as much as possible or going outside for other reasons than getting essentials (such as food and medicine). As seen in the results section, introducing incentives increased the probability of people choosing more flexible quarantine restrictions and this effect was statistically significant. Interestingly though, no one chose option 4: "Not stayed at home, visited friends or family, attended social gatherings, or travelled outside of their city". Zerbe and Paulhus (1987) define socially desirable responding (SDR) as "presenting oneself favourably regarding current social norms and standards" (Zerbe & Paulhus, 1987, p.250). In a situation such as the recent outbreak of a global pandemic, being in quarantine becomes the current social norm and standard. Following quarantine rules as strictly as possible therefore becomes a socially desirable behaviour under these conditions. There have been lots of research on people and this so-called “socially desirable responding” as a bias in surveys (Alker, 1968; Zerbe & Paulhus, 1987; Brenner, 2011; Mneimneh, Heeringa, Tourangeau & Elliott, 2014). Additional research regarding the use of truth-telling incentives on questions about socially desirable behaviours would be interesting for future investigations.

Seeing that monetary incentives using choice-matching can, in fact, reduce the SDR becomes a promising result for survey designers interested in people’s true behaviours. One particular industry that would be extremely benefited from reducing the SDR is insurance, particularly health insurance and car insurance. Insurance companies could introduce incentives in questions asking for smoking or drinking frequency and recklessness degree of driving, for example. This could help reduce information asymmetry and moral hazard between insurance companies and their potential clients.

3.4 Limitations

There are some limitations to his research that are worth mentioning. One of these is related to the sample, referring particularly to the sample size and the sample’s characteristics. Even though Cvitanić et al. (2019) argue that choice-matching is not subject to the “large sample requirement”, it would have been interesting to test if this was indeed true, by comparing a large sample and a smaller sample using choice-matching incentives. Additionally, the sample for this research is not homogeneous, including both students and non-student participants. This fact could affect the power of the statistical tests. Another limitation could be related to the number of questions and choices. In this research, multiple-choice questions were limited to three or four choices per question. It would be interesting to test truth-telling incentives for multiple-choice questions with five or more possible choices. This would, for example, allow participants to choose from a broader range the degree to which they feel a subjective feeling, eliciting thus a more accurate and refined response. However, one needs to consider time as a restriction. It took participants around 10 minutes to complete the survey. Adding more questions or more options, although ideal, would have made the survey longer and cause more participants to drop out from the survey. Considering the pandemic situation would have also more likely caused people to drop out if the survey had been longer since, as seen in the results, people were already experiencing trouble concentrating.

Vernon Smith (1982) proposed five precepts required to achieve control in an economic experiment, namely *nonsatiation, saliency, dominance, privacy, and parallelism*. The *dominance* principle refers to the reward structure, i.e. the incentives, being large enough so that people think hard enough about the questions and their answers. It is possible that the incentives used for this research did not satisfy the dominance principle. Participants in the incentives group were told that only three participants would be chosen to receive the €10 prize. This fact could cause some participants to consider that their chance of winning was very low and thus make them indifferent to the incentives. Guaranteeing monetary payments to all participants based on their score would have been better; however, budget restrictions were present. As mentioned before, it is interesting that not a single participant answered that they had not stayed home and had actually gone outside to visit friends or family or travel. Having more dominant incentives could have maybe incentivized some participants to confess to this option on the "stay home" question.

Having the chance to redo this investigation with more time and with a higher budget, some features of the experimental design could be changed and improved. First of all, collecting a larger sample with more representative participants could strengthen and generalize the results. Online services such as Amazon MTurk are very useful in these situations. By using such an online service one can specify the number of participants and, furthermore, the specific characteristics of the participants one wishes to apply the experiment to. Additionally, one can set the monetary incentives, and these could be paid automatically to all participants based on their scores. With higher financial incentives, more questions could be added to the survey and more options to the multiple-choice questions. Most of the questions in this survey had only three choices that served as a kind of rating from 1 to 3 system. Having a larger sample with a larger set of choices, that is, questions with "ratings" from 1 to 5 instead of from 1 to 3, could produce more robust results. Furthermore, more questions regarding behaviour would be an interesting addition. Finally, re-doing this survey once the coronavirus situation gets better would also be quite interesting.

Chapter Four: Conclusion

Overall it seems that truth-telling incentives do not have an effect on subjective well-being and health-related questions as long as there are no biases present. This implies that for these types of questions it is important that survey designers avoid including cognitive biases in their questions. Participants of this investigation did not underreport negative emotions regardless of the group they were assigned to, perhaps because of the harshness of the coronavirus pandemic in people's well-being at the early stage of the outbreak. Additionally, there were no statistical differences between the two groups regarding the average probability participants were willing to accept to undertake the risky medical procedure. The average probability was around *40%*, which means that people prefer to avoid a high chance of dying, even when considering living under the conditions of Health State A.

The last section of the survey included different types of questions. It was relieving, but also expected, to see that participants were handwashing more frequently than usual. Handwashing during a pandemic became a socially strongly desired behaviour and there were no differences between the control and incentives groups. There were no differences either regarding their beliefs towards the reaction of their governments to the coronavirus outbreak. The majority of participants responded that the reaction had not been sufficient. Although a subjective perception, it was expected since, at the time, confirmed cases and deaths were on the rise both in Europe and in The Americas, where the majority of the sample was from. Not presenting coronavirus symptoms was also expected since testing was not available to anyone. Surprisingly though, incentives had a statistically significant effect on the question about staying at home. Participants in the presence of truth-telling incentives were more likely to report following more flexible quarantine rules. However, incentives might have not been enough to encourage participants to confess to visiting friends or family. During a situation as the coronavirus pandemic, following strict quarantine measures becomes the social norm and the socially desired behaviour. In line with the research of John, Loewenstein, and Prelec (2012), truth-telling incentives seem to encourage people to confess socially misbehaviour.

Future research in the effectiveness of truth-telling incentives could focus on socially desired responding (SDR). Furthermore, efforts could be concentrated in developing more powerful truth-telling mechanisms that are also more comprehensible for respondents. Cvitanić, Prelec, Riley, and Tereick (2019) managed to develop a truth-telling mechanism that seems to be as powerful as the Bayesian Truth-Serum, but easier to understand by the general population. Future truth-telling mechanisms could also explore incentives other than monetary. One idea could be including a leader board where participants would see how they rank based on their prediction score. The survey designer could choose to either privately reveal individual ranking or make the leader board public.

A final note…

Conducting this investigation has taught me numerous things. It has helped me strengthen technical, analytical, and statistical skills and, more importantly, how to effectively communicate key findings. Furthermore, it has provided me with essential tools for becoming a better researcher; and I have to say, it has also ignited an internal passion for future research. It has also given me the chance to study human behaviour during a unique and unprecedented situation such as a global pandemic. Oftentimes researchers would wish to study human behaviour during such emergencies, but experimenters cannot re-create analogous drastic events. Conducting a research during these unfortunate times has been a unique and unforgettable opportunity. Overall, it has been a great experience.

Reference

Alker, H. A. (1968). Coping, Defense And Socially Desirable Responses. *Psychological reports*, 22(3), 985-988.

Baillon, A., Bleichrodt, H., & Granic, G. D. (2020). Incentives in Surveys. *Unpublished*.

Berinsky, A. J. (2018). Telling the truth about believing the lies? Evidence for the limited prevalence of expressive survey responding. *The Journal of Politics*, 80(1), 211-224.

Brenner, P. S. (2011). Investigating the biasing effect of identity in self-reports of socially desirable behavior. *Sociological Focus*, 44(1), 55-75.

Cvitanić, J., Prelec, D., Riley, B., & Tereick, B. (2019). Honesty via Choice-Matching. *American Economic Review: Insights*, 1 (2): 179-192.

De Bruin, W. B. (2011). Framing effects in surveys: How respondents make sense of the questions we ask. In Perspectives on Framing, edited by Gideon Keren, 303–325. New-York, NY. Taylor and Francis Group, *Psychology Press.*

Dolan, P., Gudex, C., Kind, P., & Williams, A. (1996). Valuing Health States: A Comparison of Methods. *Journal of Health Economics*, 15(2), 209–209. https://doi.org/10.1016/0167-6296(95)00038-0

Edwards, A. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology*, 37(2), 90-93. doi:10.1037/h0058073

EQ-5D. (nd). EuroQol. Consulted April 2020, from https://euroqol.org/

Hyman, H. (1944). Do they tell the truth?. *The Public Opinion Quarterly*, 8(4), 557-559.

Jacquemet, N., Joule, R. V., Luchini, S., & Shogren, J. F. (2013). Preference elicitation under oath. *Journal of Environmental Economics and Management*, 65(1), 110-132.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 23(5), 524-532.

Mneimneh, Z. N., Heeringa, S. G., Tourangeau, R., & Elliott, M. R. (2014). Bridging Psychometrics and Survey Methodology: Can Mixed Rasch Models Identify Socially Desirable Reporting Behavior?. *Journal of Survey Statistics and Methodology*, 2(3), 257-282.

Özdemir, S., Johnson, F. R., & Hauber, A. B. (2009). Hypothetical bias, cheap talk, and stated willingness to pay for health care. *Journal of health economics*, 28(4), 894-901.

Parry, H. J., & Crossley, H. M. (1950). Validity of responses to survey questions. *Public Opinion Quarterly*, 14(1), 61-80.

Preisendörfer, P., & Wolter, F. (2014). Who is telling the truth? A validation study on determinants of response behavior in surveys. *Public Opinion Quarterly*, 78(1), 126-146.

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science*, 306(5695), 462-466.

Smith, V. L. (1982). Microeconomic systems as an experimental science. *The American Economic Review*, 72(5), 923-955.

Van Osch, S. M., Wakker, P. P., Van Den Hout, W. B., & Stiggelbout, A. M. (2004). Correcting biases in standard gamble and time tradeoff utilities. *Medical Decision Making*, 24(5), 511-517.

Van Osch, S. M., & Stiggelbout, A. M. (2008). The construction of standard gamble utilities. *Health economics*, 17(1), 31-40.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309), 63-69.

Weaver, R., & Prelec, D. (2013). Creating truth-telling incentives with the Bayesian truth serum. *Journal of Marketing Research*, 50(3), 289-302.

Zerbe, W. J., & Paulhus, D. L. (1987). Socially desirable responding in organizational behavior: A reconception. *Academy of Management Review*, 12(2), 250-264.

Appendix

Appendix A: Survey

Figure A.1 Control group instructions

Thank you very much for participating in this survey.

Completing this survey should take approximately 8-10 minutes.

You will be answering multiple choice questions and participate in a game.
The *game* will ask you to predict the share of participants responding to each of the options of the multiple choice questions.

Example:
Question - Who do you think will win the U.S. 2020 elections?

a) Donald Trump
b) Joe Biden
c) I don't know

Game - Out of 100 participants, please estimate how many think that Donald Trump will win, how many think Joe Biden will win, and how many do not know who will win.

Your answers are important for my research, please answer truthfully.

The data will only be used for the purpose of this study and will be treated confidentially and anonymously.

Please click on the arrow on the bottom right of the page to begin.
Thank you very much in advance!

Figure A.2 Incentives group instructions (part I)

Thank you very much for participating in this survey.

Completing this survey should take approximately 8-10 minutes.

You will be answering multiple choice questions and participate in a game.
The *game* will ask you to predict the share of participants responding to each of the options of the multiple choice questions.

Example:
Question - Who do you think will win the U.S. 2020 elections?

a) Donald Trump
b) Joe Biden
c) I don't know

Game - Out of 100 participants, please estimate how many think that Donald Trump will win, how many think Joe Biden will win, and how many do not know who will win.

You will receive a **Score** based on:
(1) the accuracy of your predictions on the *game* and
(2) the accuracy of the prediction of other respondents who chose the same options as you did for the multiple choice questions

Figure A.3 Incentives group instructions (part II)

The three participants with the highest ***Score*** will be contacted via email and will receive a €10 prize.
You may choose to remain anonymous and not provide your email. In this case, you will not be eligible to win the €10 prize.

All data in this survey will only be used for the purpose of this study and will be treated confidentially and anonymously.

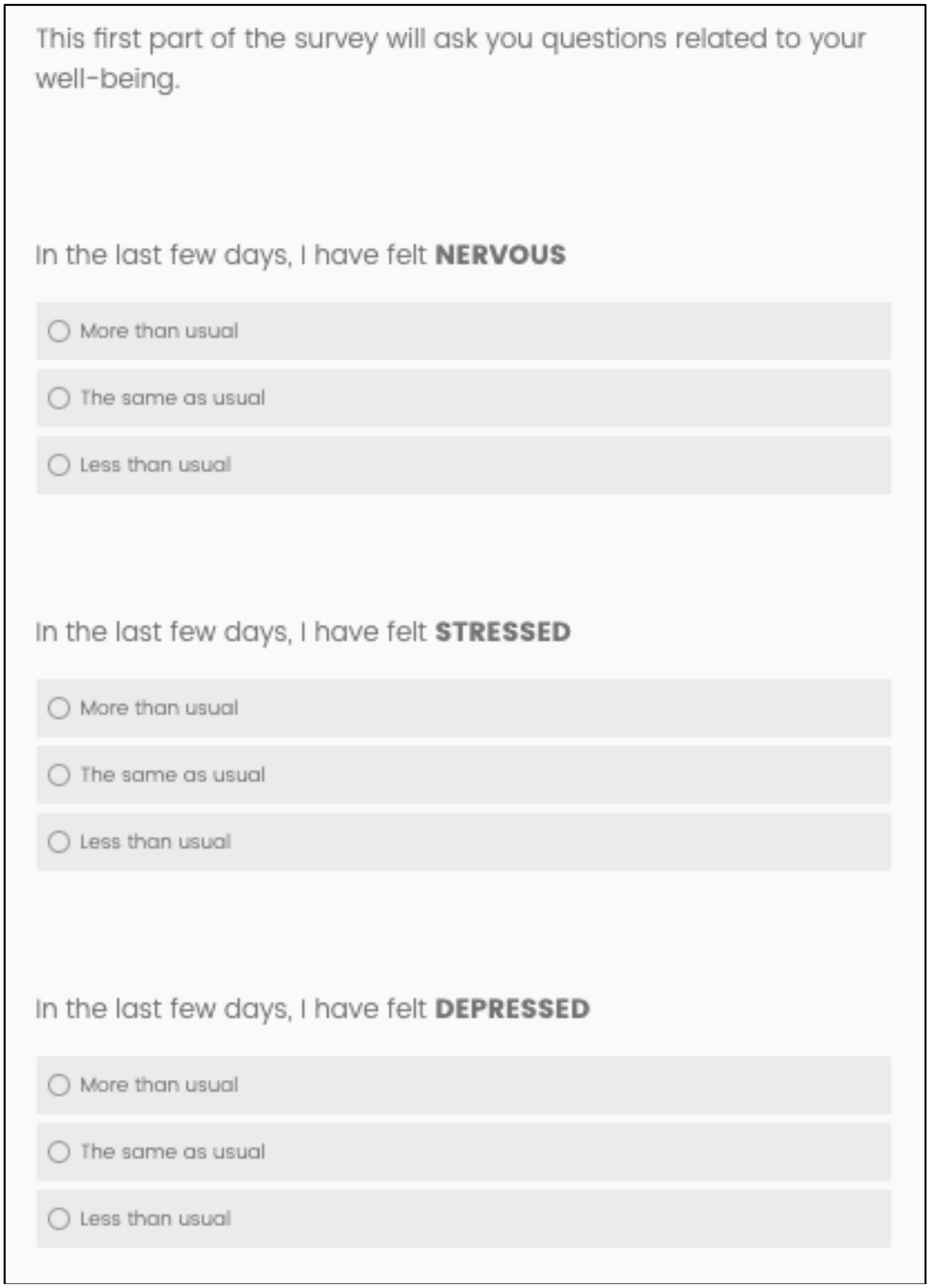
Please enter your email if you would like to have the chance to win €10 based on your *Score.*

Please note that the email will only be used for this sole purpose and will not be related to your answers.
If you wish to remain anonymous, you may leave this empty and click on the arrow at the bottom of the page to begin

Please click on the arrow at the bottom right of the page to begin.

Thank you very much in advance!

Figure A.4 Subjective well-being section



This first part of the survey will ask you questions related to your well-being.

In the last few days, I have felt **NERVOUS**

- More than usual
- The same as usual
- Less than usual

In the last few days, I have felt **STRESSED**

- More than usual
- The same as usual
- Less than usual

In the last few days, I have felt **DEPRESSED**

- More than usual
- The same as usual
- Less than usual

In the last few days, I have felt **FATIGUE**

- ○ More than usual
- ○ The same as usual
- ○ Less than usual

In the last few days, I have felt **TROUBLE CONCENTRATING**

- ○ More than usual
- ○ The same as usual
- ○ Less than usual

You will now be asked to estimate how many people, out of 100 participants, do you think would answer each possible option of the following questions

Out of 100 participants, how many people do you think would answer feeling **NERVOUS** in the last few days
(Total must add up to 100)

Number of participants feeling NERVOUS

	0 10 20 30 40 50 60 70 80 90 100	
More than usual		0
The same as usual		0
Less than usual		0
Total:		**0**

Out of 100 participants, how many people do you think would answer feeling **STRESSED** in the last few days
(Total must add up to 100)

Number of participants feeling STRESSED

	0 10 20 30 40 50 60 70 80 90 100	
More than usual		0
The same as usual		0
Less than usual		0
Total:		**0**

Out of 100 participants, how many people do you think would answer feeling **DEPRESSED** in the last few days
(Total must add up to 100)

Number of participants feeling DEPRESSED

	0 10 20 30 40 50 60 70 80 90 100	
More than usual		0
The same as usual		0
Less than usual		0
Total:		**0**

Out of 100 participants, how many people do you think would answer feeling **FATIGUE** in the last few days
(Total must add up to 100)

Number of participants feeling FATIGUE

	0 10 20 30 40 50 60 70 80 90 100	
More than usual		0
The same as usual		0
Less than usual		0
Total:		**0**

Out of 100 participants, how many people do you think would answer having **TROUBLE CONCENTRATING** in the last few days
(Total must add up to 100)

Number of participants feeling TROUBLE CONCENTRATING

	0 10 20 30 40 50 60 70 80 90 100	
More than usual		0
The same as usual		0
Less than usual		0
Total:		**0**

Figure A.5 Health section

In the next section, you will be presented with questions about health scenarios.
Please read carefully and imagine yourself in the described situation.

Please click on the arrow on the bottom right of the page to continue.

Consider the following health states:

Health State A:

- Unable to walk
- Moderate trouble with self-care activities (e.g. washing or dressing)
- Moderate problems doing usual activities (work, study, family, or leisure)
- Severe pain or discomfort
- Severely anxious or depressed

Full Health:

- No problems to walk
- No trouble with self-care activities
- No problems doing usual activities
- No pain or discomfort
- Not anxious or depressed

Imagine yourself living for the rest of your life in Health State A.

Now, imagine your doctor tells you there is a risky medical procedure that might make you live in **FULL HEALTH** for the rest of your life.

The risky medical procedure has two possible outcomes:

Probability *p* of being successful, making you live in **FULL HEALTH** for the rest of your life, and
Probability *(1 - p)* of not being successful, causing your immediate **DEATH**.

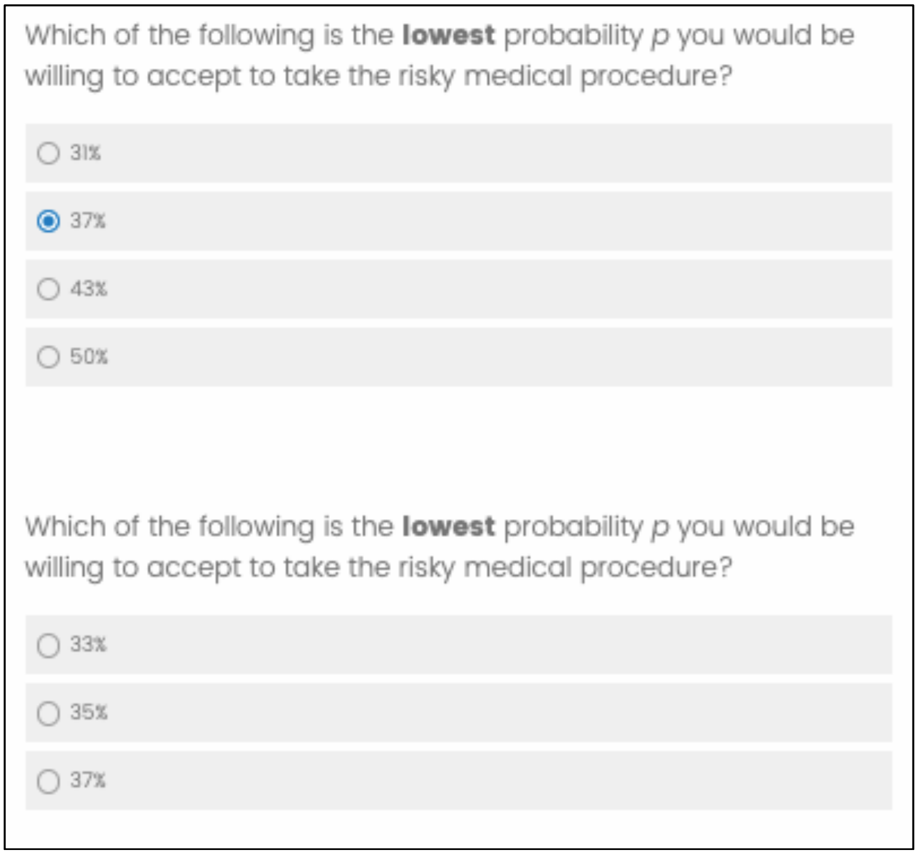
Which of the following is the **lowest** probability *p* you would be willing to accept to take the risky medical procedure?

- ○ less than or equal to 25%
- ◉ 50%
- ○ 75%
- ○ 100%

Out of 100 participants, how many people do you think would choose each one of the possible probabilities *p*?
(Total must add up to 100)

0 10 20 30 40 50 60 70 80 90 100

less than or equal to 25%	0
50%	0
75%	0
100%	0
Total:	**0**



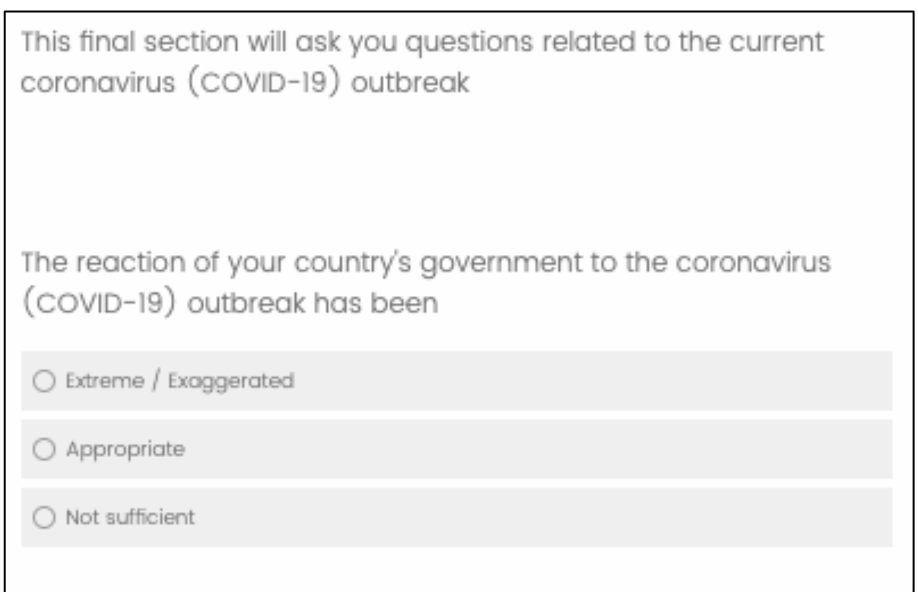
Which of the following is the **lowest** probability *p* you would be willing to accept to take the risky medical procedure?

- ○ 31%
- ◉ 37%
- ○ 43%
- ○ 50%

Which of the following is the **lowest** probability *p* you would be willing to accept to take the risky medical procedure?

- ○ 33%
- ○ 35%
- ○ 37%

Figure A.6 Behaviour and beliefs during the coronavirus outbreak section



This final section will ask you questions related to the current coronavirus (COVID-19) outbreak

The reaction of your country's government to the coronavirus (COVID-19) outbreak has been

- ○ Extreme / Exaggerated
- ○ Appropriate
- ○ Not sufficient

During the COVID-19 crisis, I have

- ○ Stayed at home
- ○ Stayed at home, except for getting essentials (e.g. food, medicine)
- ○ Stayed at home, except for getting essentials (e.g. food, medicine), going to work, doing physical activity, or walking a pet
- ○ Not stayed at home, visited friends or family, attended social gatherings, or traveled outside of my city

During the COVID-19 outbreak, I have washed my hands

- ○ More frequently than I used to before the outbreak
- ○ The same frequency as I used to before the outbreak
- ○ Less frequently than I used to before the outbreak

Have you experienced some (or all) of the coronavirus COVID-19 symptoms?

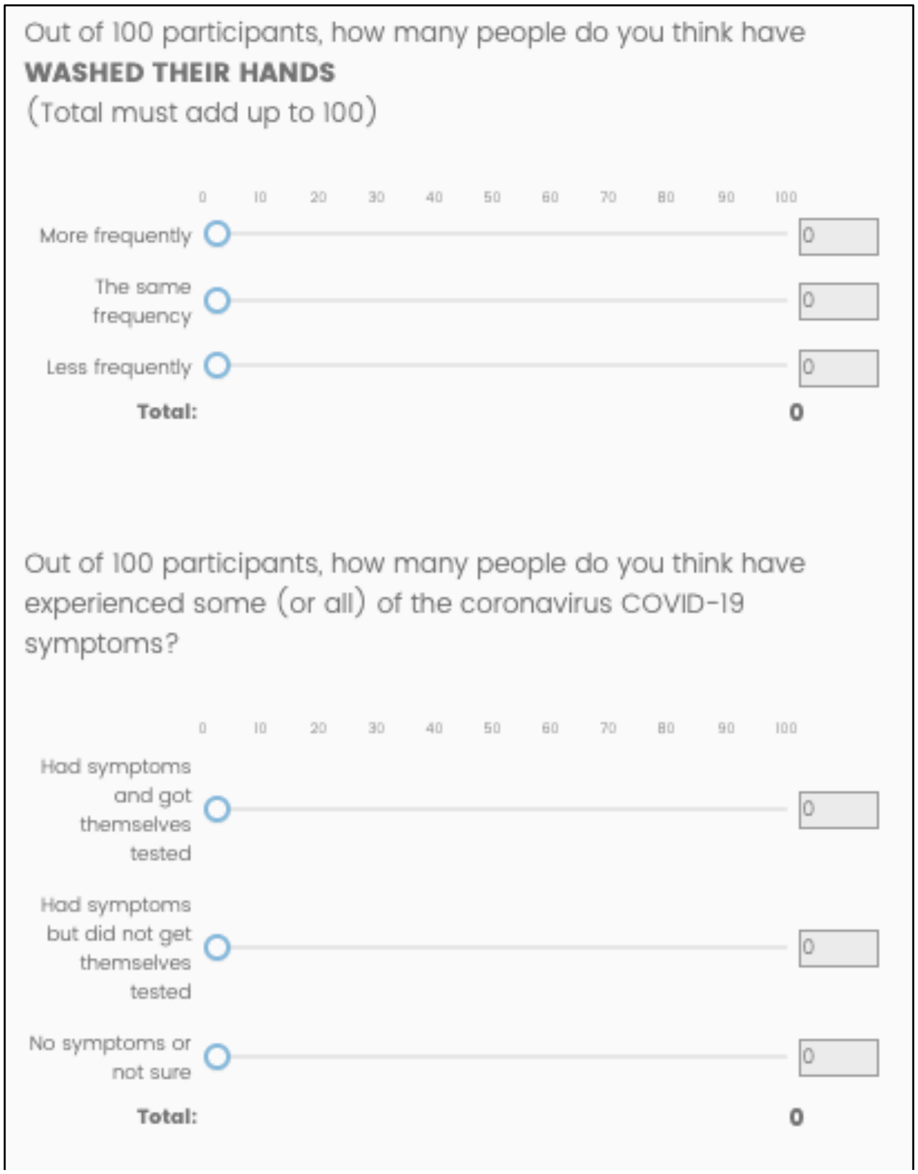
- ○ Yes, and I got myself tested
- ○ Yes, but I did not get myself tested
- ○ No or I'm not sure

Out of 100 participants, how many people do you think would answer each one of the options regarding the **REACTION OF THEIR GOVERNMENT** to the coronavirus outbreak
(Total must add up to 100)

	0 10 20 30 40 50 60 70 80 90 100	
Extreme / Exaggerated		0
Appropriate		0
Not sufficient		0
Total:		**0**

Out of 100 participants, how many people do you think have
(Total must add up to 100)

	0 10 20 30 40 50 60 70 80 90 100	
Stayed at home		0
Stayed at home, except for getting essentials (e.g. food, medicine)		0
Stayed at home, except for getting essentials (e.g. food, medicine), going to work, doing physical activity, or walking a pet		0
Not stayed at home, visited friends or family, attended social gatherings, or traveled outside of their city		0
Total:		**0**

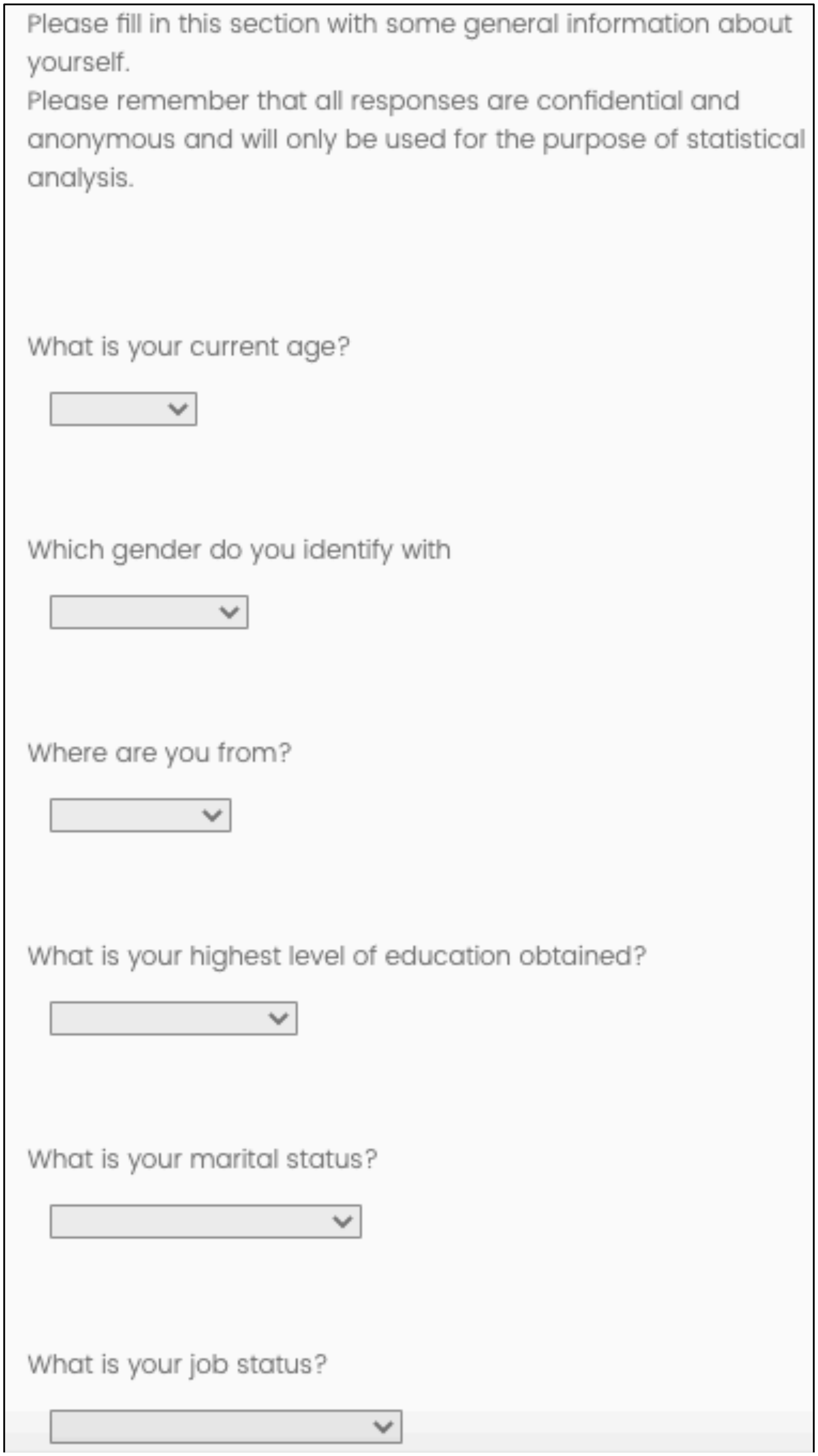

Out of 100 participants, how many people do you think have
WASHED THEIR HANDS
(Total must add up to 100)

	0	10	20	30	40	50	60	70	80	90	100	
More frequently												0
The same frequency												0
Less frequently												0
Total:												**0**

Out of 100 participants, how many people do you think have experienced some (or all) of the coronavirus COVID-19 symptoms?

	0	10	20	30	40	50	60	70	80	90	100	
Had symptoms and got themselves tested												0
Had symptoms but did not get themselves tested												0
No symptoms or not sure												0
Total:												**0**

Figure A.7 Demographic questions



Please fill in this section with some general information about yourself.
Please remember that all responses are confidential and anonymous and will only be used for the purpose of statistical analysis.

What is your current age?

Which gender do you identify with

Where are you from?

What is your highest level of education obtained?

What is your marital status?

What is your job status?

Appendix B: Variable Description

Following is the code book describing each of the variables collected and used for the analysis.

Table B.1: Code Book

Variable	**Description**
incentives	Indicator of whether participant is assigned to control group (0) or treatment group (1)
subjective_feelings	Indicator of the self-assessed well-being: less than usual (1); the same as usual (2); more than usual (3). One variable was created for each subjective feeling: nervous, stressed, depressed, fatigue, and trouble concentrating
p_i	Probability p for choosing the risky medical procedure in question $i \in \{1,2,3\}$ of the health section
gov_react	Indicator of the perceived reaction of the government to the coronavirus outbreak: not sufficient (1); appropriate (2); extreme / exaggerated (3)
stay_home	Indicator of respondent's behaviour regarding staying home during the coronavirus outbreak: Stayed at home (1); Stayed at home, except for getting essentials (2); Stayed at home, except for getting essentials, going to work, doing physical activity, or walking a pet (3); Not stayed at home, visited friends or family, attended social gatherings, or travelled outside of their city (4)
handwash	Indicator of respondent's behaviour regarding handwashing during the coronavirus outbreak: less frequently (1); the same frequency (2); more frequently (3)
covid_symptoms	Indicator of presenting COVID-19 symptoms: No or not sure (1); Yes but did not get tested (2); Yes and got tested (3)
age_cat	Categorical variable for age: Under 18 years old (1); 18-24 years old (2); 25-34 years old (3); 35-44 years old (4); 45-54 (5); 55 or older (6)
gender	Indicator of whether participant is male (1), female (2), or non-binary / not disclosed (3)
married	Indicator of whether a participant is married / co-habiting (1) or single / divorced / widowed (2)

region	Indicator of participant's region of origin: Asia (1), EMEA (2), Latin America (3), or North America (4)
education	Indicator of participant's level of education: high school (1), college (2), bachelors (3), masters (4), or PhD (5)
job	Indicator of whether participant is a student (1), employed (2), or unemployed or retired (3)

Additionally, for each of the variables from the three main sections there was a variable corresponding to the prediction (or *auxiliary game* question) of each of the options from the MCQ's.

Appendix C: Regressions

Table C.1: Linear Regression results for the relationship between score and incentives

This table presents the results of the OLS regression of Score on incentives. The reference categories used for the categorical variables age_cat, gender, region, education, and job are respectively: under18, male, LatinAmerica, bachelor, and employed.

Variables	Score
incentives	-0.126
	(0.0812)
Age group: 18-24	0.813
	(0.578)
Age group: 25-34	0.733
	(0.579)
Age group: 35-44	0.790
	(0.592)
Age group: 45-54	0.765
	(0.594)
Age group: over55	0.548
	(0.595)
female	0.184**
	(0.0767)
Other gender	0.814**
	(0.388)
Asia	-0.245
	(0.278)
EMEA	-0.293**
	(0.138)
NorthAmerica	-0.0545
	(0.115)
highschool	0.636**
	(0.284)
college	0.151
	(0.133)
master	0.330***
	(0.0977)
phd	0.199
	(0.160)

married	-0.196**
	(0.0847)
student	-0.0712
	(0.111)
unemployed	0.471***
	(0.148)
Constant	1.460**
	(0.578)
Observations	174
R-squared	0.276

Note. Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table C.2: Ordered probit regression results for the relationship between incentives and subjective feelings

This table presents the results of the ordered probit regressions. In all cases, the dependent variable is a categorical variable from 1 to 3 based on the *subjective feelings* as described in the Code Book shown in Appendix B. The reference categories used for the categorical variables age_cat, gender, region, education, and job are respectively: under18, male, LatinAmerica, bachelor, and employed.

Variable	(1)	(2)	(3)	(4)	(5)
	nervous	**stressed**	**depressed**	**fatigue**	**concentrating**
incentives	-0.0934	0.108	0.111	0.223	0.304
	(0.196)	(0.201)	(0.187)	(0.190)	(0.200)
Age group: 18-24	6.519	6.514	-5.241	5.161	7.296
	(391.4)	(211.2)	(150.7)	(218.4)	(162.6)
Age group: 25-34	6.744	7.001	-4.269	5.29	7.27
	(391.4)	(211.2)	(150.7)	(218.4)	(162.6)
Age group: 35-44	6.694	6.888	-4.465	5.008	6.981
	(391.4)	(211.2)	(150.7)	(218.4)	(162.6)
Age group: 45-54	7.109	7.33	-4.736	4.835	7.362
	(391.4)	(211.2)	(150.7)	(218.4)	(162.6)
Age group: over55	7.044	6.339	-4.765	4.652	7.023
	(391.4)	(211.2)	(150.7)	(218.4)	(162.6)

female	0.2	-0.00188	0.275	-0.0197	0.2
	(0.185)	(0.189)	(0.178)	(0.178)	(0.187)
other gender	-6.959	-2.323*	0.352	-0.238	0.349
	(269.3)	(0.984)	(0.960)	(0.958)	(1.058)
Asia	0.14	0.502	0.28	-0.434	0.0918
	(0.693)	(0.739)	(0.646)	(0.623)	(0.755)
EMEA	0.159	0.934*	0.29	0.269	-0.213
	(0.341)	(0.396)	(0.324)	(0.327)	(0.349)
NorthAmerica	0.455	-0.00986	0.0687	-0.124	0.173
	(0.290)	(0.280)	(0.268)	(0.260)	(0.281)
highschool	0.236	0.918	0.832	-0.256	1.072
	(0.843)	(1.057)	(0.712)	(0.665)	(0.964)
college	-0.109	-0.121	0.750*	0.493	0.217
	(0.323)	(0.316)	(0.318)	(0.310)	(0.330)
master	-0.402	-0.209	0.0149	0.0275	-0.0779
	(0.233)	(0.244)	(0.225)	(0.226)	(0.232)
phd	0.374	-0.0321	-0.233	-0.391	0.857*
	(0.403)	(0.388)	(0.369)	(0.379)	(0.435)
married	0.111	-0.0226	-0.183	0.282	-0.671***
	(0.208)	(0.211)	(0.196)	(0.199)	(0.204)
student	-0.032	0.028	0.136	0.073	0.365
	(0.267)	(0.271)	(0.256)	(0.261)	(0.280)
unemployed	-0.088	0.204	0.270	0.320	-0.438
	(0.358)	(0.398)	(0.348)	(0.354)	(0.348)
/					
cut1	5.231	5.71	-5.025	4.595	6.085
	(391.4)	(211.2)	(150.7)	(218.4)	(162.6)
cut2	6.859	6.737	-3.645	5.708	7.312
	(391.4)	(211.2)	(150.7)	(218.4)	(162.6)
N	174	174	174	174	174
pseudo R-sq	0.086	0.069	0.055	0.043	0.106

Note. Standard errors in parenthesis
* p<0.05, ** p<0.01, *** p<0.001

Table C.3: Linear regression results for the relationship between incentives and reported probabilities $p_1, p_2, and\ p_3$

This table presents the results of the OLS regressions. The dependent variables are the probabilities $p_1, p_2, and\ p_3$. The reference categories used for the categorical variables age_cat, gender, region, education, and job are respectively: under18, male, LatinAmerica, bachelor, and employed.

Variable	**(1)**	**(2)**	**(3)**
	p1	**p2**	**p3**
incentives	0.0636	0.0689	0.0708
	(0.040)	(0.046)	(0.047)
Age group: 18-24	0.541	0.645	0.663
	(0.285)	(0.329)	(0.336)
Age group: 25-34	0.532	0.643	0.666*
	(0.286)	(0.330)	(0.337)
Age group: 35-44	0.49	0.591	0.607
	(0.292)	(0.337)	(0.344)
Age group: 45-54	0.435	0.551	0.574
	(0.293)	(0.338)	(0.345)
Age group: over55	0.517	0.64	0.66
	(0.294)	(0.338)	(0.346)
female	0.0129	0.029	0.0302
	(0.038)	(0.044)	(0.045)
other gender	-0.236	-0.362	-0.38
	(0.192)	(0.221)	(0.226)
Asia	-0.0662	-0.0916	-0.0981
	(0.137)	(0.158)	(0.162)
EMEA	0.0278	0.0453	0.0498
	(0.068)	(0.079)	(0.081)
NorthAmerica	-0.0411	-0.0596	-0.0618

	(0.057)	(0.065)	(0.067)
highschool	0.223	0.285	0.298
	(0.140)	(0.162)	(0.165)
college	-0.0381	-0.0573	-0.0572
	(0.066)	(0.076)	(0.078)
master	-0.021	-0.00195	0.00338
	(0.048)	(0.056)	(0.057)
phd	0.119	0.073	0.0643
	(0.079)	(0.091)	(0.093)
married	-0.0085	-0.0097	-0.0174
	(0.042)	(0.048)	(0.049)
student	0.0411	0.0506	0.054
	(0.055)	(0.063)	(0.065)
unemployed	-0.104	-0.147	-0.153
	(0.073)	(0.084)	(0.086)
Constant	-0.0145	-0.216	-0.252
	(0.285)	(0.329)	(0.336)
N	174	174	174
R-sq	0.103	0.107	0.111

Note. Standard errors in parenthesis

* p<0.05, ** p<0.01, *** p<0.001

Table C.4: Ordered probit regression results for the relationship between incentives and behaviour and beliefs during the coronavirus outbreak

This table presents the results of the ordered probit regressions. In all cases, the dependent variable is a categorical variable from 1 to 3 based on the questions from the *coronavirus section* as described in the Code Book shown in Appendix B. The reference categories used for the categorical variables age_cat, gender, region, education, and job are respectively: under18, male, LatinAmerica, bachelor, and employed.

Variable	**(1)**	**(2)**	**(3)**	**(4)**
	gov_react	**handwash**	**stay_home**	**covid_symptoms**
incentives	0.186	0.106	0.497*	-0.0569

	(0.224)	(0.266)	(0.193)	(0.304)
Age group: 18-24	3.486	-4.369	5.386	4.74
	(160.9)	(870.1)	(212.6)	(376.5)
Age group: 25-34	2.72	-4.583	5.808	4.224
	(160.9)	(870.1)	(212.6)	(376.5)
Age group: 35-44	2.926	-4.661	5.769	4.001
	(160.9)	(870.1)	(212.6)	(376.5)
Age group: 45-54	3.042	0.125	5.565	4.022
	(160.9)	(925.8)	(212.6)	(376.5)
Age group: over55	2.261	-3.622	4.997	3.534
	(160.9)	(870.1)	(212.6)	(376.5)
female	-0.480*	0.445	-0.0733	0.0783
	(0.218)	(0.265)	(0.180)	(0.269)
other gender	1.517	-1.139	-0.0628	-4.877
	(1.065)	(1.131)	(0.891)	(219.6)
Asia	1.064	-0.825	0.55	0.664
	(0.698)	(0.848)	(0.656)	(0.779)
EMEA	0.979**	-0.0696	0.389	0.0103
	(0.361)	(0.483)	(0.325)	(0.484)
NorthAmerica	0.349	-0.0527	0.289	0.203
	(0.308)	(0.357)	(0.272)	(0.421)
highschool	-1.019	-0.413	0.00916	0.723
	(0.839)	(0.917)	(0.657)	(0.818)
college	-0.865*	-0.575	0.441	0.725
	(0.436)	(0.439)	(0.319)	(0.444)
master	0.184	-0.392	0.0421	0.294
	(0.260)	(0.317)	(0.229)	(0.353)
phd	0.31	0.192	0.282	1.023*
	(0.428)	(0.564)	(0.379)	(0.494)
married	0.232	0.0861	-0.174	0.0927
	(0.234)	(0.289)	(0.199)	(0.304)

student	-0.152	0.42	-0.0385	0.369
	(0.303)	(0.362)	(0.261)	(0.370)
unemployed	0.281	0.0927	-0.213	1.849***
	(0.409)	(0.524)	(0.346)	(0.441)
/				
cut1	3.472	-5.340	4.737	6.103
	(160.9)	(870.1)	(212.6)	(376.5)
cut2	5.202		6.450	7.027
	(160.9)		(212.6)	(376.5)
N	174	174	174	174
pseudo R-sq	0.159	0.093	0.067	0.182

Note. Standard errors in parenthesis

* p<0.05, ** p<0.01, *** p<0.001

Appendix D: Statistical Tests

Table D.1: Two-sample Mann-Whitney U test for subjective well-being section

Mann-Whitney U test: nervous

Group	Obs	Rank Sum	Expected
Control	107	9,501.5	9,362.5
Incentives	67	5,723.5	5,862.5
combined	174	15,225.0	15,225.0

H_0: nervous(incentives==0) = nervous(incentives==1)	
z	0.479
p-value	0.6317

Mann-Whitney U test: stressed

Group	Obs	Rank Sum	Expected
Control	107	9,297.5	9,362.5
Incentives	67	5,927.5	5,862.5
combined	174	15,225.0	15,225.0

H_0: stressed(incentives==0) = stressed(incentives==1)	
z	-0.228
p-value	0.8196

Mann-Whitney U test: depressed

Group	Obs	Rank Sum	Expected
Control	107	9,231.0	9,362.5
Incentives	67	5,994.0	5,862.5
combined	174	15,225.0	15,225.0

H_0: depressed(incentives==0) = depressed(incentives==1)	
z	-0.440
p-value	0.6597

Mann-Whitney U test: fatigue

Group	Obs	Rank Sum	Expected
Control	107	9,031.5	9,362.5
Incentives	67	6,193.5	5,862.5
combined	174	15,225.0	15,225.0

H_0: fatigue(incentives==0) = fatigue(incentives==1)	
z	-1.097
p-value	0.2725

Mann-Whitney U test: concentrating

Group	Obs	Rank Sum	Expected
Control	107	8,987.5	9,362.5
Incentives	67	6,237.5	5,862.5
combined	174	15,225.0	15,225.0

H_0: concen-ting(incentives==0) = concen-ting(incentives==1)	
z	-1.280
p-value	0.2004

Table D.2: Two-sample Mann-Whitney U test for health section

Mann-Whitney U test: p1

Group	Obs	Rank Sum	Expected
Control	107	8,870.0	9,362.5
Incentives	67	6,355.0	5,862.5
combined	174	15,225.0	15,225.0

H_0: p1(incentives==0) = p1(incentives==1)	
z	-1.593
p-value	0.1111

Mann-Whitney U test: p2

Group	Obs	Rank Sum	Expected
Control	107	8,951.5	9,362.5
Incentives	67	6,273.5	5,862.5
combined	174	15,225.0	15,225.0

H_0: p2(incentives==0) = p2(incentives==1)	
z	-1.278
p-value	0.2014

Mann-Whitney U test: p3

Group	Obs	Rank Sum	Expected
Control	107	8,945.0	9,362.5
Incentives	67	6,280.0	5,862.5
combined	174	15,225.0	15,225.0

H_0: p3(incentives==0) = p3(incentives==1)	
z	-1.294
p-value	0.1956

Table D.3: Two-sample Mann-Whitney U test for behaviour and beliefs during the coronavirus outbreak section

Mann-Whitney U test: gov_react

Group	Obs	Rank Sum	Expected
Control	107	9,231.5	9,362.5
Incentives	67	5,993.5	5,862.5
combined	174	15,225.0	15,225.0

H_0: gov_react(incentives==0) = gov_react(incentives==1)	
z	-0.503
p-value	0.6152

Mann-Whitney U test: handwash

Group	Obs	Rank Sum	Expected
Control	107	9,328.0	9,362.5
Incentives	67	5,897.0	5,862.5
combined	174	15,225.0	15,225.0

H_0: handwash(incentives==0) = handwash(incentives==1)	
z	-0.170
p-value	0.8694

Mann-Whitney U test: stay_home

Group	Obs	Rank Sum	Expected
Control	107	8,577.5	9,362.5
Incentives	67	6,647.5	5,862.5
combined	174	15,225.0	15,225.0

H_0: stay_home(incentives==0) = stay_home(incentives==1)	
z	-2.713
p-value	0.0067

Mann-Whitney U test: covid_symptoms

Group	Obs	Rank Sum	Expected
Control	107	9,554.5	9,362.5
Incentives	67	5,670.5	5,862.5
combined	174	15,225.0	15,225.0

H_0: covid_symptoms(incentives==0) = covid_symptoms(incentive	
z	0.992
p-value	0.3211