Erasmus University Rotterdam

Erasmus School of Economics
MSc Behavioural Economics

Self-deception and the trustworthiness of political polling

Author:
B. Odenhoven

Supervisor:
Prof. dr. H. Bleichrodt

11th of May 2018

Abstract

This paper studies to what extent self-deception is related to someone's reported voting intentions. Specifically, it studies whether extreme party voters are more prone to self-deceptive behaviour than non-extreme party voters. This finding could serve as an explanation for the large discrepancy found in election polling. Especially as mixed-method pre-election polling is becoming more and more used, it is likely the discrepancy can be explained by the behaviour of people rather than the method of polling. To test this, an online experiment is conducted. The experiment that is used is for a great extent based on the experiment of Mijović-Prelec and Prelec (2010), who use it to find evidence for self-deception under students. The experiment makes use of Korean characters, which participants must classify as either male or female. Using a logit model to analyse the answer patterns, there is no significant evidence to conclude that being an extreme-party voter influences the likelihood of showing self-deceptive behaviour. However, overall a large percentage of self-deceptive answers is found, showing that this psychological tendency is indeed present.

Keywords: self-deception, self-signalling, diagnostic utility, election polling,

Acknowledgement

I want to express my appreciation and gratitude to my supervisor Han Bleichrodt, for supporting and motivating me in the journey of writing this thesis. After writing a short paper about it in the seminar ‘Applied Behavioural Economics’, Han Bleichrodt encouraged me to continue exploring the topic. Since I started a job whilst writing this thesis, it must have taken a great deal of patience for Han to guide me along the process of finishing it. His optimism and positivity were big motivating factors.

Moreover, I want to thank Mijović-Prelec, Prelec and Ursa Bernardic for inspiring me with the experiment they conducted for the paper ‘Self-deception as self-signaling: a model and experimental evidence’, and for sharing the data of their findings. The topic of self-deception and diagnostic utility is something that interests me a lot, and something that hasn’t been studied to a large extent in the field of Behavioral Economics. As I think self-deception influences many decisions besides the reporting of voting intentions and therefore affects the way that we should predict human behavior, I hope this will be explored more in the future.

Lastly, I want to thank my parents for their patience and for exerting the right amount of pressure that was needed for me to write this thesis.

Contents

1. Introduction

When it turned out that Trump had won the elections in the U.S.A. on the 6th of November 2016 a worldwide shock prevailed, despite the fact that 90% of the pre-election polls predicted Clinton to win (Henderson, 2016). As argued by Crespi (1988), interest in pre-election polls has been based on the expectation that these polls can provide accurate indications of the election outcomes, which was not the case in the given example. More so, the Trump case is not the only example of an inaccurate prediction, as over time there have been many cases where the pre-election poll results were not in line with the actual outcome of the elections. Examples of this include the presidential race between Reagan and Carter in 1980 and the outcome of the Brexit referendum in the UK in 2016. The pre-election polls three weeks before the election in 1980 predicted a race that was too close to call; however, Reagan won the actual election with ten points difference. Moreover, 114 of the 272, only 42%, of the polls that were held in the UK from the moment the Brexit referendum was announced, predicted the actual outcome correctly: that the UK would exit the EU. These cases are a few examples of inaccurate poll predictions, and as argued by Goeree and Grosser (2007), pre-election polls often get it wrong.

Quite some research has been done on the effects of polling methods and the accuracy of the poll-predictions. These methods can differ from telephone surveys to door-to-door surveys. Different methods cause different biases in the outcomes, and it is therefore generally advised to implement a mixed-strategy method. Besides that, research on misreporting voting intentions focuses on the psychological biases or theories that come into play when people are asked to state their voting preferences. As mixed-strategy methods have become increasingly popular and polling-method biases can no longer explain the wrong predictions, this research will focus mainly on the misreporting of voting intentions as a reason for inaccurate predictions.

In the previous literature, behavioral economical theories have merely been applied to explain the discrepancy between the percentage of people that reports to vote, and the actual voter turnout. Examples of such research are the papers of both Struthers and Young (1989) and Sears and Funk (1990), who both investigate the effects of self-interest and altruism by applying various classical and behavioral economic theories to the decision-making process of voting or abstaining. One field that has not been extensively studied is the presence of psychological biases in the decision-making process of answering pre-election surveys. Research has found that in the past, when telephone surveys and real-life surveys were the main polling methods, there has been a type of interviewer-effect, in which people report to be more socially desirable than they actually intend to vote (Streb et al., 2007). Now that mixed-strategy methods, including online survey methods, are being used, it is possible that there is some kind of 'self-effect', in which respondents don't respond socially desirable for the sake of another person, but for the sake of themselves. This could be due to two reasons, either the person truly believes he/she is more socially desirable than he/she is, or he/she wants to send a message of having favorable traits to him- or herself, to receive a temporary satisfaction from doing so. One 'self-effect' that can be used to explain this effect is self-deception.

Self-deception is mainly studied in the field of psychology and can be simply explained as the process of lying to oneself, or the biased information flow within an individual (Trivers, 2000). Self-deception is closely related to self-signaling, which can be described as partly choosing an action to secure good news about one's traits or abilities (Bodner and Prelec, 2003). The explanation of self-deception as a form of self-signaling and the relation to reported voting-intentions will be further explored in this thesis. More specifically, this thesis will test whether 'extreme' party voters are more likely to be self-deceptive when compared to non-extreme voters. When it is shown that this is indeed the case, this could explain inaccurate pre-election polling predictions. The research in this thesis is inspired by Mijović-Prelec and Prelec (2010), who used a sophisticated experiment to prove that self-deception is a form of self-signaling. This experiment will be repeated and adjusted, so that it fits the goal of this research.

This thesis will be organized as follows: first, all relevant background information is summarized to give a better understanding of the topics in this thesis. This first section will provide information on polling methods as well as on the concepts self-signaling and self-deception. Second, the hypothesis is formulated which will later be tested with statistical tests. Third, the methodology will describe both the experiment and the formal model that is used to explain the intuition behind this experiment. Fourth, the data gathered from the experiment and the results of the statistical tests will be presented, and the paper will finish with the discussion and conclusion.

2. Literature review

As stated by Hillygus (2011), who explored the evolution of polling in the United States, the Gallup's quota-selected polls in 1936 marked the beginning of scientific election polls. The Gallup poll, also called the American Institute of Public Opinion, contacted respondents either by mail or by personal interviews to find out who they were planning to vote for. Gallup believed that relying on chance to select polling participants was not sufficient to create a sample that would accurately resemble the whole population (Berinksy, 2006). For this reason, Gallup created quotas for their interviewers, so that a representative sample would be interviewed. Under these quotas, the researcher targeted predetermined proportions of people from specific segments of the population, which were based on gender, age, economic class and occupation. As argued by Hogan (1997), the polls created by Gallup predicted both the elections in 1940 and 1944 correctly, and after this it became America's best known and most trusted pollster. The success lasted until 1948, when Gallup predicted that Dewey would win the presidential race from Truman with something between five to fifteen percentage points. The actual outcome of the election was a victory of Truman by more than four percentage points. According to Berinsky (2006), the quota-controlled sampling procedures had led to inaccurate predictions because they did not create samples that were representative of the total population. Argued by Hogan (1997), in the years after this, Gallup took a lesson out of what happened in 1948. He changed the way he dealt with undecided voters, he found a way to screen out non-voters and changed the time the polls would be conducted so they were now conducted within a few days of the real election.

In 2016, Gallup announced that it would stop predicting presidential election outcomes, as in 2010 and 2012 its polls were far off compared to other polling institutions (Shepard, 2015). However, this didn't mark the end of pre-election polls, as since the start and evolution of the Gallup poll, election polling has evolved remarkably. As estimated by Traugott (2005), the number of trial heat polls held in the US has increased with 900 percent from 1984 to 2000. This increase comes with new and innovative methods of polling, as well as an increased number of independent parties and people conducting polls. As in earlier years polls were typically conducted using live interviews or telephone surveys, nowadays a substantial number of polls make use of 'interactive voice response' (IVR) or are held online. With the IVR-method, respondents hear a recorded automated voice, asking them which candidate they would vote for, were the election to take place today (Blumenthal, 2005). Online polls ask randomly selected participants if they want to participate in the election survey, after which the results are collected and analysed. With these simpler methods of polling, not only media organizations (such as news stations or newspapers) and political candidates perform polls, but entrepreneurial pollsters and independent polling firms nowadays also conduct them. These entrepreneurial pollsters conduct polls with a motive of profits or increased and positive publicity.

As mentioned before, pre-election polls often get it wrong. Explanations for the gap between the pre-election predictions and the actual election outcomes found in the literature can be grouped in two broad categories: those that suggest polling methodology is wrong and those that suggest that voters might misreport their voting intentions. The following paragraphs will summarize those two categories of explanations.

2.1 Polling methodology

As stated by Hillygus (2011), pollsters must decide upon a huge variety of different design decisions when it comes to designing the most accurate poll possible. As there are many choices to make about the design and the methodology of the polls, certain decisions can lead to certain biases in the outcomes. The following paragraphs will explain the most important biases that are found in poll-results.

2.1.1 Non-response bias

According to Siemiatycki and Campbell (1984), the non-response bias can be explained as the number of people that do not respond to the poll when they are being contacted. This can occur because people are not at home, they have registered wrong telephone numbers or because people refuse to answer. The non-response bias occurs mostly in pre-election political telephone polls, as the number of call-backs is relatively lower compared to other types or surveys. When trying to predict the winner of the election using the poll-results, the non-responses must be treated in a certain way. As formulated by Siemiatycki and Campbell, the problem in trying to assess the nonresponse bias is that the population from which these non-responding voters come is unknown. Also, it is difficult to assess the characteristics of the non-respondents, as this group of people will generally not answer follow-up calls. This raises the question of what to do with the non-responses; simply disregard them or divide them equally or non-equally over the running candidates. Siemiatycki and Campbell try to measure the extent of the bias in the research they conducted and use data for both the respondents and the non-respondents. They find that bad telephone numbers and not-at-homes have different bias effects than refusals. As a solution, they find that more call-backs (after the first missed call) reduces the bias associated with not-at-homes, but not the bias with bad numbers and refusals.

2.1.2 Coverage bias

As found by Blumberg and Luke (2007), in the year 2006, 32% of low-income young adults lived in households without a landline telephone; instead they owned wireless telephones such as mobile phones. As some pre-election polls make use of landline telephone surveys, this raises the question of how representable the respondents are of the real population. Argued by Busse and Fuchs (2009), the exclusion from households without a landline in the pre-election surveys leads to a coverage bias; only certain people from the population are covered by the polls. On the other hand, Keeter (2006) finds that the polls conducted by telephone performed well in 2004, even though they omitted families without landline telephones. He argues that non-coverage is a relatively small source of the total survey error, as people without telephones are less likely to vote and thus their omission won't have a big effect on the prediction. On the other hand, the author states that the percentage of households without landline is growing fast, which means that the survey error caused by the problem is likely to increase over time. This is confirmed by Mokrzycki et al. (2009), who find that the proportion of cell-only eligible voters rose from 7% in 2004 to 20% in the elections of 2008. Moreover, the authors find evidence that the coverage error will grow over time, and therefore advise researchers to make sure that they cover the cell-only population in their pre-election surveys well enough. One possible way to solve this problem is by combining different methods of polling, such as face-to-face, telephone, internet, fax and mail in a way that they rule out the disadvantages of using each method alone (Terhanian, 2008).

2.1.3 Effects of methodological problems on recent polls

As mentioned before, polling techniques have developed considerably since they were introduced decades ago. Instead of just using face-to-face interviews and telephone surveys, polling institutions are now using a variety of methods to make the most accurate predictions possible. For example, ESOMAR (European Society for Opinion and Market Research) has 4900 members in 130 different countries, which include many polling organizations. Together with 500 individuals from academic and business professions in several countries ESOMAR proposes certain guidelines for opinion polls and public surveys. In their guidelines, it is advised to use mixed methods of polling, including face-to-face interviews, telephone polls and Internet polls. As they state, the use of multiple methods within a single poll is becoming common, as it insures coverage for different groups that might be harder to reach by using one main method. This is confirmed by de Leeuw (2005), who finds that mixed-mode survey approaches give an opportunity to compensate for the weakness of each individual mode. In this way, using multiple methods ensures that the coverage bias in survey results can be reduced and maybe even removed. She also finds that mixed-mode strategies help to reduce the non-response bias, as mixed-mode strategies make it possible to start with a less costly method and in cases of high non-response rates change to a costlier method with obtaining responses.

When looking specifically at the Netherlands, there are five main political polling companies; I&O Research, Ipsos, Kantar Public, Peil.nl and De Stemming. Most of these companies use an online method to interview their respondents, however the technique of selection differs per company. Whilst in the poll of I&O Research the respondents select themselves to participate, in the poll of Kantar Public the respondents are randomly selected. Moreover, some companies invite the respondents for face-to-face interviews next to the online survey. Peilingwijzer adds together all the predictions of the five polling companies. As stated in their methodology, the website uses a special statistical model to rule out the so-called 'house-effects' of each individual polling company. Those house-effects are the differences in the panel of respondents of the firms, which can lead to a systematic bias in the results.

It would be expected that the polling-firms, by using different techniques and selections of respondents, would produce results that are quite close to the actual outcomes. However, when looking at the recent elections held in March 2017, the Peillingwijzer predicted 26 seats for the party VVD, whilst in the actual election they won 33 seats. Also, in the 2012 elections, all five polling companies predicted a big win for the SP, where the win was predicted to be between 20 and 22 seats. However, in the actual election the SP won only 15 seats. This calls for a different explanation for the wrong predictions, which have to do rather with reports of people instead of the methods used in polls. The following paragraph will explain an alternative view, which states that poll predictions are wrong because of misreported voting intentions.

2.2 Misreported voting intentions

2.2.1‘Conscious misreporting’

Burke and Taylor (2008) find that people can misreport their preferences in order to influence the voting behaviour of others. Namely, by misreporting the candidate they will vote for (reporting they will vote A whilst in reality they will vote B) they increase the turnout amongst voters that have the same preference as he/she has. This is the case as there is a certain ‘free-riding effect’ under which a person is less likely to vote the more people in the country share his preference, as voting is costly. By misreporting his/her own preferences, the person decreases the amount of people sharing the same preference, and hereby decreasing the free-riding effects in his/her own party.

In addition, Jowell et al. (1993) find the ‘shame factor’ to be another explanation why polls might be off in predicting elections. With this ‘shame factor’ people lie to the pollsters because they are ashamed to tell their true voting intentions. This effect can be generalized to a ‘social desirability bias’, in which people tend to state they support the social desirable party, even though they don’t support this party in reality. The social desirability bias is caused by a human tendency to present oneself in the best light possible, and to state opinions and answers that agree with the general view of what is correct or socially acceptable (Fisher, 1993). There has been extensive literature on this bias, and its presence is found in almost all types of self-report measures across many fields, such as economics, psychology and social science. Hence, because of social-desirable answers, pre-election polls could be wrong in predicting the winner.

2.2.2 ‘Unconscious misreporting’

In the previous paragraph, certain reasons were presented which cause people to deliberately misreport their voting intentions. With the biases described above, people know they are falsely answering the pre-election polls. However, it can also be argued that this misreporting happens unintentionally, so that people are not aware at the time of answering the poll that they will vote differently than stated. The following paragraph will explain the most important concepts that can be applied to this ‘unconscious’ misreporting.

2.2.3 Self-deception

Self-deception is a much-discussed topic in both psychology and philosophy. Different definitions and explanations have been introduced, all aiming to explain the relative complex concept. For example, Quattrone and Tversky (1984), explain the concept of self-deception by applying it to theories on decision-making. As they argue, decisions can have a good or a bad outcome, depending on which state of nature occurs. It is therefore important for the decision-maker to weigh the possible outcomes of a chosen action by the probability of the states that have an influence on the outcomes. For example, someone can make the decision to take the car to a holiday destination; which can result in a safe trip or in a car accident. The states of nature in this case are the probability of a car crash and the probability that this crash won't happen (and thus a safe trip will be made). In this example, the probability of a car crash is independent of the action of choosing a car as means of transportation. Put differently, the relevant states of nature are independent of one's choice. However, this is not always the case. Consider the situation in which somebody must choose to take the car or to walk whilst being intoxicated from alcohol. The person has to decide whether he/she wants to be home faster (by taking the car) but risking a car accident. In this example, the relevant states of nature (car accident or no car accident) are not independent of the person's choice, as driving whilst being intoxicated from alcohol increases the risk of a car accident.

Although the correlation in the previous example is quite clear, Quattrone and Tversky argue that in reality it can be difficult to recognize that some acts may be causal or diagnostic of outcomes with which they are correlated. The authors state that although people know that an action is only diagnostic of a certain outcome (and in no way causal), they still select an action because of the correlation between the action and outcome. The authors prove the following example: imagine that students can choose to attend a review right before the exam of a certain subject they're taking. It is known that the students who attend the review session get better grades on average than students who don't attend this session. However, does one last review session really help students to understand the material better and score a higher grade (causal), or does it mean that mostly the ambitious and conscious students who would do well no matter what attend this review session? The authors claim that even if students know that attending a review session is not causal of a good grade on the exam, they might still go to the review session, as attending the session is somehow related to getting a good grade (in a diagnostic manner). This happens because choosing certain actions (attending the review sessions) allow people to make positive interpretations about their own characteristics. Choosing to attend the review session might be a sign of dedication, or willingness to study (even if the student knows that he will not study enough to get a good grade). However, when looking rationally, it is implausible to assume that attending one last review session will make a student pass the course, when this student was already highly likely to fail the course, as he/she did not follow any other sessions. This might seem like a paradox; a person chooses a certain action in order to show favourable news about his/her characteristic, although he/she knows that choosing this action will not have any effect on the final outcome. The question that is raised by this paradox is: how could people make positive implications of their own behaviour, when they know that they deliberately chose this behaviour?

The authors pose two answers to this question. First, it is possible that people are unaware that their actions are not causal of the outcome, so the students would attend the session and truly believe that this will increase chances of a good grade. However, as stated before, it's not plausible to believe that attending one last session will make a student pass instead of fail, when this student was already highly likely to fail in the first place. As argued by the authors, this belief is in line with Bem's (1972) theory, which states that inferences about the self are based only on the observations of one's own behaviour. Second, it is possible that a certain level of self-deception helps people to accept the diagnosis that is implied by the chosen actions. Put differently, a certain level of self-deception can cause that group of students that knows they will most likely fail to make themselves believe that attending the review session is not just an act chosen to show dedication/willingness to study.

Gur and Sackheim (1979) explain self-deception as the state of a person in which "deep inside" an individual holds a belief that is not in line with the affirmed belief. Moreover, they say that the self-deceived person, who is holding two contradicting beliefs, doesn't do this on purpose or out of disregard, but he/she simply isn't aware of one of the beliefs. Following this line of thought, Gur and Sackheim created the following four criteria that are necessary for self-deception to take place:

1. The person holds two contradicting beliefs (belief 'p' and belief 'not p')
2. These contradicting beliefs are held at the same time
3. The person is not aware of holding one of the beliefs
4. The act that determines which belief is, and which belief is not subject to awareness is a motivated act.

Mijović-Prelec and Prelec (2010) use this theory to explain self-deception in a slightly different manner and argue that self-deception can be seen as a form of self-signalling. They state that the reasoning of Gur and Sackheim raises two paradoxes, a static paradox and a dynamic paradox. The static paradox relates to the first criteria; how can a person hold two incompatible beliefs at the same time? The dynamic paradox relates to the third and fourth criteria, namely how a person becomes self-deceived; how can a person intentionally obtain a belief and still remain unaware of this belief? As the authors state, recognizing that you are obtaining or ignoring certain beliefs seems to defeat the purpose of the effort of obtaining or ignoring this belief. To explain these two paradoxes that the theory of Gur and Sackheim brings, the authors argue that self-deception occurs because of one psychological mechanism: self-signalling.

2.2.4 Self-signalling

Bodner and Prelec (1997) describe self-signalling as choosing a certain action to provide a signal to ourselves which affects our esteem. Sometimes these actions that are chosen have no causal impact on someone's traits or skills, but people still seem to have some kind of motivation to pursue this action. For example, think of a person taking the bike to work instead of the car, even when it's raining. This choice might send a signal to the person that he/she is acting healthy and being fit, as this choice reveals positive information on those 'fitness' traits of the person.

Quattrone and Tversky (1984) conducted an experiment in which the self-signalling phenomenon was very visible. To summarize, they asked participants to put their hand in cold water until they could no longer tolerate the pain. After this first test, the participants were told that recent studies had found that the time somebody could tolerate the coldness on their hands was related to some kind of medical condition, an inborn heart disease. One group of the participants was told that increased tolerance was a sign of the heart disease, while the other group was told the opposite; a decreased tolerance related to the heart disease. The participants were then asked to perform the test again. Although the tolerance had no casual impact on whether someone had the disease or not (this was inborn), most participants changed their tolerance in the direction of the 'good news' when they were told that decreased tolerance was a sign of the disease, as they could now keep their hand longer in the cold water. This is a clear example of self-signalling, as people choose a certain tolerance level that reveals positive information on one of their traits (in this case their health).

2.2.5 Self-deception as self-signalling; the model

Mijović-Prelec and Prelec (2010) combine the two previous theories and suggest that self-deception can be seen as a case of self-signalling. To explain this argumentation, they use aspects of Bayesian game theory. This modelling technology assumes a rational agent who is defined by his/her preferences with a utility function, by beliefs that are represented by subjective probabilities and by an action or choice set that is available. When the agent has to choose between different actions, he/she is assumed to choose the action that maximizes expected utility (EU). Whenever the agent receives new information, this information is incorporated into his/her beliefs according to Bayes' rule. Bayesian game theory is used to model strategic interactions when different agents make choices, and the solution to those interactions is the Nash Equilibrium.

It is possible to add some kind of psychological structure to this Bayesian model, by dividing the agent into different entities that are simultaneously interacting with each other. These entities can be called 'selves' or 'modules'. One of those 'selves' is then responsible for choosing the action while the other of the 'selves' interprets this action. In this way, it is possible for someone to pick a certain action (only to give out a signal) and at the same time to interpret this action as being diagnostic of a good outcome. Using the same example as before, one can deliberately pick the bike to work (to try to give out a signal) and still interpret this as a sign of being fit or healthy. In other words, the assumption of the multiple selves solves the 'static paradox' of the classic theory of self-deception of Gur and Sackheim (1979); that a person is not able to hold two incompatible beliefs at the same time.

To explain the dynamic paradox, we first must elaborate on the utility model that is used to model self-signalling and self- deception.

Bodner and Prelec (1997) introduce a utility model to formulate the value of the self-signalling and its impact on someone's utility. They distinguish between outcome utility (the utility of the causal consequences of the choice), and the diagnostic utility (the value of how the action changed someone's perception of the disposition that it affected). They argue that people maximize a combination of the two sources of utility. Using the previous example, choosing the bike gives somebody an outcome utility (utility of actually performing the bike-ride) and some diagnostic utility (the utility of the self-image regarding someone's health or fitness-level that has changed after the person chose the bike instead of the car).

To formulate the model, we need to specify certain variables. Self-signalling relies on an underlying characteristic that is a) personally important, b) introspectively inaccessible and c) potentially revealed through actions (Mijović-Prelec and Prelec, 2010). This characteristic is described by θ, which can for example represent someone's fitness-level or health. This characteristic must satisfy the three criteria that were listed before. θ^0 represents the 'real' value of this characteristic that is introspectively inaccessible, so someone's actual fitness level or health. The possible action chosen by the agent is represented by 'x', which is for example the bike ride. The utility that is generated by the outcome 'x' without any choice on action is represented by $u(x,\theta)$. The $u(x,\theta)$ using the previous example is the outcome utility generated by the bike ride in case the agent does not have a car or any other options of transportation. The value of θ is unknown to the agent, and the inferences that are made on its value are represented by $p(\theta)$; the current self-image with respect to this characteristic. The value of this self-image is imposed by another function; $v(\theta)$, which represents the emotional reaction (how much pleasure or pain) of finding out the true θ. As an agent chooses a certain action over other actions, this reveals something about his/her inaccessible characteristics. Thus, by choosing an action 'x', the self-image of the person gets updated from $p(\theta)$ to $p(\theta|x)$. As stated before, this change in self-image causes 'diagnostic motivation', which can be formulated as:

(1) $$\sum_\theta v(\theta)p(\theta|x) - \sum_\theta v(\theta)p(\theta).$$

Put into words, this can be described as the difference between the emotional state of having the new self-image and the emotional state of having the old self-image, before choosing a certain action. Hence, the diagnostic utility represents the extent to which the action 'x' provides positive or negative information about the characteristic, leading to a positive and negative value respectively.

Using the example from before again, in which θ represents someone's fitness level, $u(x,\theta)$ represents the pleasure or pain associated with 'x' minutes of biking when there is no alternative option available (in case the car is broken) and $v(\theta)$ to the happiness or sadness when someone would find out his/her real fitness level. The total utility of choosing the bike to work would then be the outcome utility summed with the diagnostic utility, where λ represents the weight of the diagnostic utility:

(2) $$V(x,\ \theta^0) = u(x,\ \theta^0) + \lambda(\sum_\theta v(\theta)p(\theta|x) - \sum_\theta v(\theta)p(\theta)).$$

Formula 2 represents the model used for self-signalling by Bodner and Prelec (2003), which can be adjusted in order to include self-deception. Note that the outcome utility includes θ^0, which is the case as when there is no other option available, and the person will choose 'x' minutes of biking, he/she will receive the utility that is related to his/her actual fitness-level. Without any other option available, the biking is a natural choice, and there are no changes in the perception of the fitness-level. Hence, in this formula θ^0 is used, to distinguish between the 'perceived' value of fitness-level; θ.

With self-deception there is a deep belief that decides whether the person is self-deceptive or not, for example the deep belief that someone is healthy. To tailor the example to the topic of this research, let's consider a new example. Assume the characteristic θ represents racism, whether somebody is characterized by having inner racists thoughts or opinions. In the event of self-deception, the person may notice certain signals that can be interpreted as signs of having racist opinions, however the person remains unaware of how to interpret these signals. An example of such a signal can be that the person notices a certain uncomfortable feeling when he/she sees someone dressed in traditional Muslim clothing. This person stays unaware whether this uncomfortableness is because he/she just doesn't like the look of the clothes, or whether the clothes remind him/her of Islam or Muslims, something that prompts anger and discomfort within the person. With self-deception, the interpretation of these signals (are they signs of being racist or a random discomfort with a certain style or way of dressing) is the inaccessible characteristic represented by θ. Continuing with this model for self-deception, θ_A represents the probability of event A and $u(x, \theta)$ an expectation over these events. As θ represents the inaccessible interpretation of a certain signal, θ_A is the probability that the signal is diagnostic of event A. For example, the event A can be the case that the person is actually racist, and θ_A is the probability that this event occurs. Furthermore, $u(x, \theta)$ is the expectation of the utility when event 'A' takes occurs. This expectation can be formulated as:

$$(3) \qquad u(x, \theta) = \sum_A \theta_A U(x, A)$$

where $U(x, A)$ is the utility of an action 'x' (for example, the person telling his/her friends about his/her voting preferences) if the event A (actually being racist) occurs. Imagine that the person from the example actually intends to vote for an extreme-right party. The interpretation of the aforementioned signal could be θ_A, in which the person interprets his-/herself as actually being racist. Probably, in this situation the person will explain her extreme-right voting preferences by admitting his/her racist views. However, the interpretation could also be θ_B, in which the person might tell him-/herself that he/she is not racist, but that he/she simply dislikes the traditional Muslim clothing style. In this situation, the person could explain the extreme-right voting preferences by different reasons, such as saying that the extreme-right party will be beneficial to the current economic state of the country.

The equation of total utility is slightly different than before, as we now add the effects of self-deception. The equation becomes:

$$(4) \qquad V(x, \theta^0) = u(x, \theta^0) + \lambda \left(\sum_\theta u(x, \theta)\, p(\theta|x) - \sum_\theta u(x, \theta) p(\theta)\right).$$

In the self-signalling equation stated before in equation 1, the diagnostic utility was dependent on the change in the self-image combined with $v(\theta)$, the emotional reaction (how much pleasure or pain) of finding out the true θ. Note that in the case of self-deception, in formula 4, $v(\theta)$ is replaced by $u(x, \theta)$, which depicts the expectation of a certain event to occur. Explained differently, in the case of self-deception the total utility exists of the outcome utility and the diagnostic utility, the same as in the case of self-signalling. However, with self-deception the diagnostic utility is constructed of the change in self-image, in light of the interpretation and the corresponding event (does event 'A' or event 'B' take place?). For example, intending to vote for an extreme-right party will cause a shift in someone's self-image, and the utility that this generates depends on whether the person interprets him-/herself as a racist (event A) or not (event B). This is the case because θ now represents the interpretation of certain signals rather than the value of a specific characteristic.

Evaluating formula 4, total utility can be described as the summation of the outcome utility and the diagnostic utility. Imagine that the person observes the same signal: an uncomfortable feeling with traditional religious clothing. This can either be a sign of actual distaste, in which the person doesn't like the clothes, or this can have a different explanation (for example negative associations with the religion itself). The outcome utility is represented the same as in the self-signalling case and represents the utility of action 'x', stating the specific voting intentions, when there is no other party available; the agent has no other choice than 'x'. The interpretation of diagnostic utility changes slightly and now represents the utility of changing the self-image by choosing 'x' (the voting intention for a specific party) in light of how the person evaluates a certain signal; the expectation of having racist thoughts.

2.2.6 Static and dynamic paradoxes

With the model explained above, the static and the dynamic paradox posed by the definition of Gur and Sackheim (1979) can be solved. As stated before, assuming the agent contains of multiple 'selves' that are acting at the same time can solve the static model. More formally, these multiple selves can be denoted by θ^0; which represents the 'deep belief', x; which represents the stated belief or expressed belief and $p(\theta|x)$; which is the experienced belief, the emotional state or change in self-image after the action 'x' has been chosen.

As mentioned before, the dynamic paradox is concerned with the process of being self-deceived, specifically with the question of how a person can intentionally obtain a belief or remain unaware of a belief. Mijović-Prelec and Prelec (2010) provide two different scenarios of self-signalling to deal with these questions. In the first case, which the authors call the 'face-value' situation, the person acts without any 'diagnostic motivation', and is motivated by only the first part of total utility, the outcome utility. In other words, the person does not consider that he/she is being self-deceptive. The new self-image '$p(\theta|x)$' is composed by choosing an action 'x' that exposes a certain characteristic which maximizes the outcome utility $u(x, \theta^0)$. For example, using the earlier example, a person can choose to take the bike to work (x), which would expose the physical shape of the person that in turn maximizes a person's health-state. In this example the person truly believes that taking the bike will maximize his/her health, which results in a positive change in self-image and thus a positive experienced belief.

In the second case, which the authors call the ‘rational’ situation, the person is fully aware of the fact that he/she is self-deceptive. Looking back at the formula for total utility, a person in the ‘rational’ situation tries to maximize a combination of both outcome and diagnostic utility. In this case, the new self-image $p(\theta|x)$ is formed by the assumption that actions are chosen in order to generate the change in self-image that flows from these actions. In this way, the signalling value of the chosen action is discounted, as the person deliberately chooses the action in order to change his/her self-image.

2.3 Conscious or unconscious misreporting?

As argued before, in the Netherlands most pre-election polls are done online, if not in combination with another method. Furthermore, in the US and in the UK, online polls are also the most widely used method for pre-election polling. Hence, the misreporting of voting intentions is probably not a manifestation of the ‘shame factor’ effect, as online polls are anonymous and no face-to-face or voice-contact is required. Looking at the recent examples of polling failures, it is therefore likely that another mechanism is responsible for the misreported voting intentions. As explained, self-deception causes people to choose an action that is not in line with their deep belief. In this way, self-deception could be linked to political polling, in which the overt statement of support for a certain party does not match the ‘real’ or deep political conviction. For example, people might not want to admit to themselves that they truly support a more right-wing extreme party, which many people view as being racist, and therefore misreport their voting intentions to other people and to themselves. In this way it might be the case that some people plan and state to vote for a socially desirable party and believe that this is causal to the fact that they are perceived as a non-racist, unprejudiced person. However, this relation might only be diagnostic, as in the voting-booth on the day of the election the person will act according to his/her deep beliefs, and vote for the party that he/she truly supports. If in this case the deep-belief turns out to be different than the stated belief, the person is thought to have acted self-deceptive, in the way that the person stated/planned to act differently on the election day (to order to change the self-image) than he/she will actually do. The actual support of the ‘extreme-right’ party, and the belief that somebody will vote for a socially desirable party are in a way two incompatible beliefs, and the person is unaware one on those beliefs until some time before or on the day of actual voting.

There can be two cases, as mentioned before, of this so-called self-deception, the ‘face-value case’ and the ‘rational case’. The face-value case takes place when the person really believes that he/she will vote for the party that he/she reports in the polls, however on the day of the election the deep belief wins, and the person changes his mind when casting his/her vote. In this way, the person is completely unaware of the two incompatible beliefs until the actual day of the election. The other case occurs when the person, sometime before the election starts doubting whether he/she will really vote for the party that he/she reported in the pre-election polls, but stating this ‘socially desirable’ choice changes one self-image which gives the person some level of diagnostic motivation. At this point self-deception takes effect, as the person is aware that he/she is carrying to incompatible beliefs, and one of those beliefs is chosen to provide a change in the self-image. Since the person is aware of the two incompatible beliefs, the signalling value of the action (misreporting the voting-intentions to a more socially-desirable party) is discounted. Eventually, it does not matter which of the two cases takes place, as both cases can lead to a prediction that isn't a trustworthy reflection of the actual outcome. The following paragraph will phrase the hypotheses that will be tested in this research.

3. Hypothesis

When judging the results from the literature as discussed before, it is not reasonable to accrue the discrepancy between political-poll results and election outcomes to methodological issues only. Therefore, this research will test whether 'extreme' party voters are more likely to be self-deceptive when compared to non-extreme voters. Whether someone is self-deceptive or not will be evaluated by conducting an experiment, which is to a large extent based on the experiment of Mijović-Prelec and Prelec (2010). By linking self-deception to voting-preferences, this research will test an alternative explanation for the discrepancy in pre-election poll results. With respect to voting-preferences, the 'extreme' and 'non-extreme' parties will be classified according to how socially desirable and accepted their agenda points are. The following hypothesis will be tested to answer this question

H1: Extreme party voters are more self-deceptive than non-extreme party voters.

The following paragraphs will explain and evaluate the experiment that is done to test the abovementioned hypothesis.

4. Methodology

To test the hypothesis mentioned before, an experiment will be conducted. The experiment used in this research is based on the experiment of Mijović-Prelec and Prelec (2010), with which they prove that self-deception is a special case of self-signaling. The original experiment will not be amended too much; only some necessary changes are applied in order to fit the goal and purpose of this research.

The experiment consists of a pre-experiment (pre-phase) and the actual experiment, which consists of two parts (phase 1 and phase 2). In the pre-phase, a group of students (n=134) is asked to classify 25 Korean characters as either male or female. This pre-phase is an online experiment, and prior to the classifications, the participants are told to use their full intuition and attention throughout the whole experiment. In reality, the Korean language consists of an alphabet, and characters are formed by using different letters. In that way, there are actually no male or female Korean characters. To select participants that are unaware of this, so that they answer the questions with full sincerity, participants with prior knowledge about the Korean language are excluded from the experiment. Only three students indicated that this was the case, so the final number of eligible participants is 131. This group of participants will then see 25 different characters one by one, and after seeing each character they will state whether they think the character is male or female.

After the participants have answered all the questions, the answers can be used as a 'consensus-list' or answer-key. In this respect, twelve characters are chosen that have the highest conformity of the gender amongst all participants. The twelve characters together with the percentage of conformity can be found in the appendix. The percentages of conformity range from 81% to 90%, meaning that the characters that are chosen clearly to have something in common that people recognize as male or female characteristics. Six of those characters are classified with this high conformity as male, and the other six are classified as female. These twelve characters will be used in phase 1 and phase 2 of the actual experiment.

4.1 Actual experiment

As stated before, the actual experiment consists of phases 1 and 2, which start right after each other. In these phases, the twelve characters with the highest conformity from the pre-experiment are used. The two phases are both online experiments, and like the pre-experiment, participants with any prior knowledge of the written or spoken Korean language are excluded from the experiment. Moreover, participants from the pre-experiment are not able to participate again in the actual experiment, as they might recognize the characters and have biased answers. Phase 1 is similar to the pre-experiment, in the sense that twelve Korean characters are shown to the participants, who then must state whether they think the characters are male or female. However, in addition to the classification (M/F), the participants also must state how confident they are about their classification on a scale of 1-5. This confidence rating is asked right after they have classified the character. Moreover, in phase 1 the participants can earn points for each 'correct' classification, which will be transferred to phase 2 of this experiment. In total, 12 points can be earned, that is 1 point for each correct classification. The participants are told in the beginning of the experiment that the 'correct' answers are based on the answer-key that is determined in a pre-experiment, and thus reflects the majority-opinion of a group of previously tested subjects. The participants are told in phase 1 that the five people with the most points at the end of the experiment will win a €10 bol.com voucher.

At the end of the twelve classifications and confidence ratings in phase 1, phase 2 of the experiment starts. Phase 2 is very similar to phase 1, in the sense that the participants must classify the same twelve characters as in phase 1, which have now changed in order. Also, after each classification they are asked again for the confidence rating of the classification that they made. However, prior to the classification the participants are asked for an 'anticipation', which is a guess about whether the next character that will be shown will be a male or female character. This is a pure random guess, as the participants don't know which character will be shown and the order of the characters is random. After the anticipation, the character is shown, and the classification and confidence rating are asked. It is not possible to go back to the previous answer, and thus to change answers. In this phase, the participants can win both points for correct classifications and for correct anticipations. For both correct answers, 1 point is assigned, meaning that one can earn a total of 24 points (12 correct classifications and 12 correct anticipations).

Before this second phase of the experiment starts, the participants get a pop-up message with either a 'classification bonus' or an 'anticipation bonus'. The two different messages are randomly assigned to the participants. In the 'classification bonus', the participants are told that the three participants with the highest classification-score in phase 2 will earn a bonus of 5 extra points (on top of the points earned in phase 1 and at the end of phase 2). In the 'anticipation bonus', the participants see the same message but now the three participants with the highest anticipation-score will earn the bonus. Table 1 shows the points that can be assigned in both phases of the experiment. Both a and c are worth 1 point. For example, in Phase 2 when the character is actually male, and the participants correctly anticipates a male character (Anticip=M) and afterwards classifies it as male (Class_2=M), he/she will receive a+c, which is 2. In the experiment, the points function as the incentive, as the 5 participants with the highest amount of points will receive a bol.com voucher.

At the end of the second phase, the participants are asked which party they voted for in the general Dutch elections, held in March 2017. It is assumed that the participants will answer this question with full honesty, as they can choose to answer the experiment in anonymity. Only when they want to participate in winning the bol.com vouchers they will have to fill in their email-address, but they can also choose to leave this field blank. Also, as stated in the literature review, the social desirability bias is reduced in Internet surveys, as participants are not affected by the influence of hearing or seeing the interviewer. It is therefore plausible to assume that the participants will answer in full honesty to this question.

Table 1. Possible rewards in the experiment, with a=c=1

	Phase 1		**Phase 2**			
Character	Class$_1$= M	Class$_1$= F	Anticip= M	Anticip= F	Class$_2$= M	Class$_2$= F
M	c	0	a	0	c	0
F	0	c	0	a	0	c

4.1.1 Intuition behind the experiment

As explained by Mijović-Prelec and Prelec (2010), self-signaling entails that when people are driven by receiving good news, they will have biased temporary or short-run decisions, even when those decisions have a negative effect on the long-run goal. As the authors illustrate, imagine a person with zero experience in sport that decides to become fitter. When this person is driven by good news, he might go to the gym every day for two hours, as this makes him feel good and proud about himself (provides good news). However, for an untrained person this can have serious effects on the long run, as the gym visit should be build up gradually in order to become fitter without getting injured. The authors state that this 'self-signaling' bias should also be present in tasks which are unusual to the participant, and in which the incentives are merely financial, without being directly related to the self-image or self-esteem of the participant. When this is the case, the authors argue that one should be able to generate self-deception with stimuli and incentives in a repeated and reliable manner.

In the experiment, the tasks of classifying and anticipating 12 characters are small and arbitrary. The tasks are explained well and are easy to understand, so that the participants will know what is expected from them. The incentives are purely financial, as 'winning' the most points will reward five participants with a voucher. The reward-structure in the experiment is responsible for triggering self-deceptive behavior from the participants, as it rewards both the anticipation as well as the classification. For example, in the situation that someone anticipated a female character to show up, but the actual character that is shown is male (as the participant classified it as male in phase 1), the participant can either choose to confirm the anticipation (and receive a) or to acknowledge the wrong anticipation (and receive c). A self-deceptive person would therefore classify a character as female in phase 1, but after anticipating a male character, he/she would classify this same character as male in phase 2. According to Mijović-Prelec and Prelec (2010), the subject may wish to believe the character is actually male, in order to approve the prior anticipation. As the rewards 'a' and 'c' are the same, both 1 point, there is actually no financial motivation to be self-deceptive, the benefit is purely psychological. This is the case as confirming the anticipation will increase the rewards with 'a', but at the same time decrease the rewards with 'c', as the classification in phase 2 is wrong.

In order to increase the motivation for self-deception, the two different bonus systems are used. With the 'anticipation bonus', the participants have an increased psychological incentive to be self-deceptive. This happens as under this bonus-system, the participant receives extra points when he/she performs relatively well in anticipating the correct characters, giving a psychological incentive to confirm one's anticipations. However, confirming one's 'wrong' anticipation, and hereby providing self-deceptive classifications does not lead to higher financial incentives. Contrarily, these self-deceptive classifications probably lead to less correct classifications, hereby reducing the amount of points that are received. The 'anticipation bonus' therefore only triggers psychological benefits, but at the same time it poses real financial costs (the reduction in classification-points).

4.1.2 Political parties

The aim of this research is to find out whether voters of extreme parties are more self-deceptive than voters of non-extreme parties. To gather data on voting preferences, the participants are asked for which party they voted in the latest Dutch elections, held in March 2017. From all parties that participated in this election, we have to distinguish 'extreme' from 'non-extreme' parties. Although opinions on this matter are probably varied, we will use the table introduced by Marks et al. (2002) in order to make the distinction between extreme and non-extreme parties. Marks et al. investigated whether party positioning is a predictor for a parties' position on the issue of European integration. In order to do so, they grouped parties into so-called 'party families', based on their programmatic commitments, views on economic integration and their views on European political integration. On one side of this spectrum there are the extreme left parties, characterized by left, and somewhat communist positions on markets, welfare, social justice and democratic decision making. These extreme left parties strongly oppose European economic and political integration, which is believed to increase economic inequality and to decrease the capacity to regulate markets.

On the other side of the spectrum are the extreme right parties. These parties are characterized by strong programmatic commitments about national culture, national sovereignty and national defense. Additionally, these parties oppose European economic and political integration, which is believed to undermine national control.

Beside the parties that can easily be placed on the political spectrum, there are parties in the Netherlands for which this is more difficult. Such parties are characterized as single-issue parties, which are supported predominantly on the basis of one single issue. An example of a single-issue party is Artikel 1, started in 2016 with the aim to reduce racism, discrimination and social injustice. Such single-issues will also be counted as 'extreme' parties, as they often deal with controversial issues in which they take an extreme stand. As most single-issues parties are small and get a relatively small number of votes, the effect of this inclusion is believed to be relatively small. Table 2 provides an overview of the Dutch parties that will be considered as 'extreme' in this research. Together with the name of these parties, table 2 shows whether they are considered extreme right, left or as a single-issue party. Moreover, the table shows how many seats the party won in the elections of March 2017, to show the support for the parties. In the Netherlands, there is a total of 150 seats available in the House of Representatives. A complete overview of all political parties can be found in Appendix B.

Table 2. Dutch extreme parties by type and size

Party name	Left/right/single-issue party	Nr. of seats
PVV	Right	20
SGP	Right/orthodox Calvanist	3
Forum voor Democratie	right	2
SP	Left	14
Denk	Left	3
Artikel 1	Single-issue	0
GeenPeil	Single-issue	0
OndernemersPartij	Single-issue	0

4.2 Model

In order to formulate the psychological benefit that is obtained by acting self-deceptive, we can use the model of self-deception as form of self-signaling (formula 4) as explained before.

Applying formula 4 to this experiment, θ^0 represents the 'real' belief about the character, so the 'deep belief' whether the character is male (M) or female (F). This 'deep belief' can be obtained by looking at the answer of the participant that was given in phase 1 of the experiment, when there was no incentive for self-deceptive behavior. As mentioned before, in the case of self-deception, it is the interpretation of certain signals that is represented by the 'inaccessible characteristic', depicted by θ. In phase 2 of the experiment, it is the anticipation-answer that acts as the signal that needs to be interpreted. The participants have to choose whether to confirm this signal (by classifying in accordance with the anticipation) or to disconfirm this signal (by classifying differently than the anticipation). The probability of a character to actually be male can be denoted as θ_M and the probability that the character is female is denoted as θ_F, those are the so-called 'events' that can take place. In this respect, the agent chooses a certain action 'x', which is the classification in phase 2 (Class$_2$), with which the participant either confirms or disconfirms the anticipation. The action 'x' can either be a male-classification, in which x=m, or a female classification in which x=f. The diagnostic utility depends on the expectation of the deep belief (θ^0) in light of the action that is chosen. In other words, when a participant chooses to confirm or disconfirm his/her anticipation, what does this say about his/her deep belief of the character? Imagine a situation in which the subject anticipated the next character to be male. The two utilities that are generated by choosing either of the two actions are

$$x = m: \qquad \mathbf{(5)}\ V(x = m,\, \theta^0) \ = \ (a + c)\, \theta^0_m \ + \ 0 \ \ \theta^0_F \ + \ \lambda\, ((a + c)\, E(\theta_M \,|\, x = m) \ + \ 0\, E(\theta_M \,|\, x = f)$$

$$x = f: \qquad \mathbf{(6)}\ V(x = f,\, \theta^0) \ = \ a \ \ \theta^0_M + \ c \ \ \theta^0_F \ + \ \lambda(a\, E(\theta_M |\, x = f\ ) \ + \ cE(\theta_F | x = f)) + \ \ 0\, E(\theta_M \,|\, x = f)$$

In both cases, either event θ_M or event θ_F takes place, depending on whether a male or a female character is shown. This event takes place randomly and cannot be influenced by the participant. In both utility-formulas, $\theta^0_{M/F}$ stands for the actual deep belief, a belief that is inaccessible for the participant. Looking at the first situation (5), where the subject classifies the character as male, it can be seen that his total utility V partially consists of the outcome utility; $(a + c)\ \theta_M^0$+ $0\ \theta_F^0$. As stated before, the outcome utility is the utility that arrives from choosing a certain action when there is no alternative available, so when there is no other choice than to classify as male (in equation 5) or as female (in equation 6). The first part of the outcome utility; $(a + c)\ \ \theta_M^0$ occurs when the character is male, which is then correctly anticipated and classified by the person and for which he/she receives a+c. If the character shown were to be female, θ_F^0 , the participant would earn nothing (0), which is shown by the second part; $0\ \ \theta_F^0$.

In the second situation (6), where the participant chooses $x = f$, the outcome utility again depends on whether the character shown is male or female, denoted as θ_M^0 or θ_F^0 respectively. In this first case, with θ_M^0, the participant will give the correct anticipation and receives 'a', whilst in the second case the participant will give the correct classification and receive 'c'.

Now, to predict the right course of action, we can subtract formula 5 from 6. This will formulate the difference in utility of choosing $x = m$ over $x = f$. In the absence of diagnostic utility, so in the case of $\lambda = 0$, the subject will classify the character as male only when θ_M^0 > θ_F^0, as the reward 'a' is equal to the reward 'c'. This happens only when the probability of a male-character is bigger than 0.5. The actual probabilities of θ_M^0 and θ_F^0, are 0.5, because of the fact that the characters are ordered randomly and the chances of a male or female character to show up are equal. Note that the participants are not told whether their answers are correct after phase 1. Hence, they remain unaware in phase 2 whether the male/female classifications made in phase 1 are correct. For this reason, it could be assumed that not all participants recall exactly how often they answered male/female in the first phase, when going into the second phase. If they do recall this, it is possible that the classifications also provide some kind of outcome utility.

However, as can be seen from both equations, the total utility also depends on the diagnostic utility, as long as $\lambda \neq 0$. As mentioned before, the diagnostic utility depends on the expectation of the deep belief, denoted as $E(\theta_{M/F}|\ x = m/f)$. Although the person classifies the character as female, $x = f$, the deep belief of the character can either be θ_M^0 or θ_F^0, and it is the expectation of this deep belief that drives diagnostic utility. When classification happens symmetrically (the subject is not biased towards a classification), $E(\theta_M\ |\ x = m)$ is equal to $E(\theta_F|\ x = f)$. With this simplification, it is clear when someone would classify a character as male, namely when equation 5 is bigger than equation 6; $V(x = m, \theta^0)$ - $V(x = f, \theta^0)$> 0;

$$c\ (\ \theta_m^0\ -\ \theta_F^0\ ) +\ \lambda a(E(\theta_M|\ x = m\ )\ -\ E(\theta_M|\ x = f)) > 0$$

Again, when self-signaling is absent ($\lambda = 0$), the subject will maximize the first part of the equation, which means that he/she will choose $x = m$ only when θ_M^0 > θ_F^0. However, when self-signaling is present ($\lambda \neq 0$), the diagnostic motivation has to be taken into account as well. When looking at the formula above, the diagnostic motivation can be formulated as $E(\theta_M\ |\ x = m)$ - $E(\theta_M\ |\ x = f)$. Put in words, the diagnostic motivation can be seen as the difference in expectation of the deep belief to be male, when someone chooses $x = m$ or when someone chooses $x = f$. That is, before giving the anticipation and the second classification, the participants already provided the first-classification, where there was no incentive to provide a self-deceptive answer. The first classification can be seen as the 'deep-belief' of the person, which is the true belief about the character (male or female). In the second phase of part 2, the order of the character changes, and so the participant might not exactly recall or access the deep belief about the specific character. Diagnostic utility measures the difference in what someone expects his/her deep belief to be, inferred by classifying it for the second time as male $(x = m)$ or female $(x = f)$. Is someone convinced that his/her deep belief is really male, by providing a second classification as male? Or does this person know that his/her deep-belief is actually different from their second classification, but confirming the anticipation provides an extra psychological spur? Put another way, the diagnostic utility measures the extent to which self-deception gives people some kind of extra utility.

This diagnostic motivation can be measured by taking the differences in the first and second confidence rating, as argued by Mijović-Prelec and Prelec (2010). As they state, using this difference excludes the variation in intrinsic confidence that each person has when it comes to the classifications, and at the same time it removes variation of how the rating-scale is used.

4.2.1 Self-deception in the model

The actual deep-belief of a person can be obtained by looking at someone's first classification (Class_1), when no incentive for self-deception is present. Self-deception enters the formula when the deep-belief of the character is not in line with the anticipation, and the person chooses to confirm the anticipation anyway. For example, the person classified the character in phase 1 as male, but in phase 2 he/she anticipates a female character to be shown. When the person is asked to classify this character again, he/she changes his/her answer, and now instead classifies the character as being female. The possible patterns of answers are depicted in table 3, with their corresponding answer-type. In this table, the first letter in the pattern represents classification 1, the middle letter represents the anticipation and the latter represents the last (or second) classification. When someone has a consistent answer, he/she answers the same (M=male, F=female) for all three questions. Someone is acting inconsistent, when he/she answers the first classification with for example male, then correctly anticipates male, but answers female in the last classification. The person disconfirms their own anticipation and 'deep-belief', and hereby acts inconsistent. Moreover, a person answers 'honest', when he/she disconfirms their anticipation but confirms their 'deep-belief', depicted by the first classification. On the contrary, a self-deceptive answer confirms the anticipation, but hereby disconfirms the 'deep-belief'. This last pattern, the self-deceptive answer type, is the one that is of interest in this research.

Table 3. Answer-types and their corresponding pattern

Pattern	Type
MMM/FFF	Consistent
MMF/FFM	Inconsistent
MFM/FMF	Honest
MFF/FMM	Self-deceptive

4.2.2 Statistical model

In order to find evidence for the hypothesis, the data collected from the experiment will be analyzed. In total, 329 people participated in the online experiment. Note that this number includes everyone that even opened the experiment. In the results section it will be analyzed how many of these people answered all questions, and of which participants the data can actually be used in the analysis. The participants were gathered through different online channels, such as Facebook and Twitter. In order to get the most diverse selection possible, the experiment was posted to websites and pages from different political parties.

To analyze whether voting preferences are related to self-deception, a regression will be run. All participants will be put into one of the two voting groups; extreme-voters or non-extreme voters. Hence, our data-set can be viewed as a panel-data set. The following statistical model will be used to analyze the data,

$$C_{i,j}^2 = C_{i,j}^1 + A_{i,j} + Group_i * A_{i,j} + e_{i,j}$$

In words, the dependent variable is C^2, which denotes the second classification of person *i* at character *j*. The character *j* stands for the answer-patterns provided for each of the 12 characters. In this sense, time j=1 corresponds to the first character, j=12 to the last. The independent variables consist of the first classification, C^1, the anticipation, *A,* and the interaction term $Group * A$. The $Group$ variable depicts whether somebody belongs to the extreme-voters group ($Group = 2$) or to the non-extreme voters group ($Group = 1$). As the data on the dependent variables is dichotomous (binary), and the data on the other variables is categorical, a logit model is the applicable model to use. The model above can be transformed into a logit model, which is formulated as:

$$logit(C_{i,j}^2 = 1) = C_{i,j}^1 + A_{i,j} + Group_i * A_{i,j} + e_{i,j}$$

In this logit model the dependent variable can be described as the probability that the second classification is answered as male/female, based on the anticipation, the first classification and the interaction effect of the anticipation answer and the voting-group. In other words, how likely is someone to provide a second classification as male, when the anticipation is male, and he/she is an extreme-party voter? The estimation of the coefficient of the anticipation ($A_{i,j}$) will show whether people are self-deceptive. Besides that, the estimation of the coefficient of the interaction term will show whether people of one voting group are more self-deceptive than the other group.

As Stock and Watson (2007) argue, estimating a model with a dependent variable that is dichotomous and binary, using a linear probability model causes heteroskedasticity. This means that the error terms have a non-constant variable across changing values of the independent variables. Using a logit model solves this problem. Additionally, a common problem in 'normal' panel-data sets is serial correlation amongst the error terms. Serial correlation can be explained as the correlation of the error terms over different time-periods. This serial correlation occurs in time-series when the error term of one period carries information into future time periods (Pindyck & Rubinfeld, 1988). Serial correlation doesn't cause the estimated coefficients to be biased or inconsistent, but it could influence their efficiency.

Since the data used is in panel-data format either a random effects or a fixed effects model can be used to estimate the variables. In a random effects model, the variation across people is assumed not to be correlated with the independent variables (Torres-Reyna, 2007). In other words, random effects assume that the individual error term of each person is not correlated with the independent variables. In a fixed effects model, this assumption is relaxed. With using a fixed effects model, the characteristics of each person are allowed to be correlated with the independent variables in the model. When judging the model described above, it is likely that characteristics are correlated with the variables. For example, someone's education or voting-preferences of their parents may influence whether that person belongs to the 'extreme-voters' or not. Hence, a fixed effects model will be used. A Hausman test will be performed to check whether the appropriate model to use is indeed a random effects model. As mentioned by Torres-Reyna (2007), this test shows whether the unique errors of each person are correlated with the independent variables.

5. Results

In total, 329 people participated in the online experiment, of which 226 finished all questions. The other 103 people stopped in the middle of the experiment or finished as soon as they were asked for which party they voted. One possible reason for this high number of people not finishing the experiment is the total length of the experiment, consisting of two phases. Another possible reason could be the reluctance of people to state which party they support in the second part of the experiment, although the experiment was answered anonymously. The platform that is used for the questionnaire; Qualtrics, estimated that answering all questions would take approximately 12 minutes. However, the average answering time of the 226 participants who finished the experiment is found to be 900,5 seconds, which is a little bit longer than 15 minutes.

From the 226 participants, 119 were shown the anticipation-bonus message, and 107 participants were shown the classification-bonus. These bonus-messages were randomly shown to the participants. 57 participants answered anonymously, without leaving their email-address, whilst the other 169 chose to leave their email address to compete in winning the vouchers.

Figure 1 shows the distribution of voting preferences amongst the participants. The party 'VVD' is the biggest, with 41% of the participants that voted for this party. Second comes D66, with 21%. From the participants, nobody voted 50Plus, so this party is not visible in the chart. Under 'Anders' are participants that voted differently than the parties listed. Examples of this are the single-issue parties GeenPeil and Artikel 1, but also participants that were too young to vote or that abstained from voting. In the actual elections of 2017, VVD was indeed the biggest party. However, the PVV was the second party that got most of the votes, something that isn't represented in this study.

Figure 1. Distribution of voting preferences

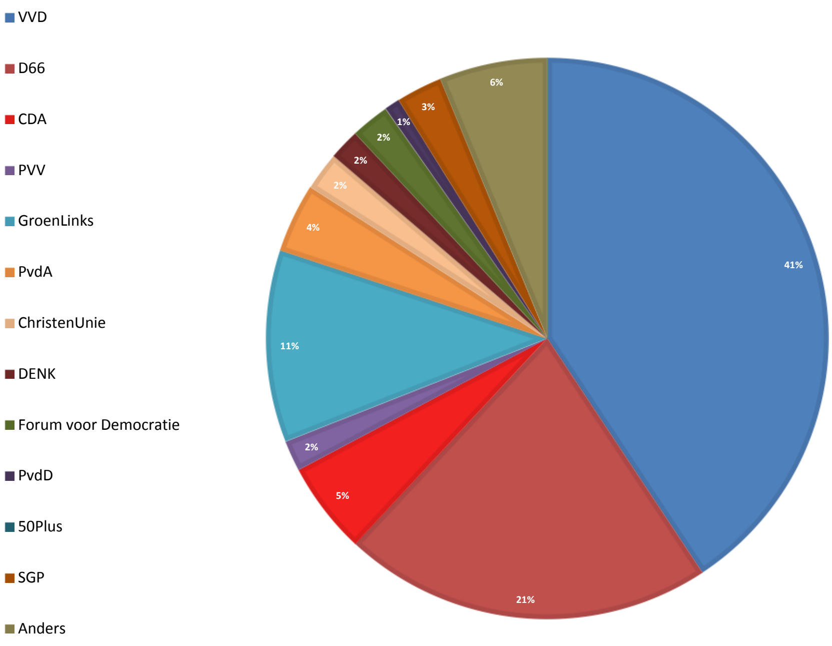


Table 4. Distribution of the answer-patterns sorted by bonus-group

	confirming pattern (C2=A)		Disconfirming patterns (C2≠ A)	
	consistent	self-deceptive	inconsistent	honest
classification-bonus	27%	35%	22%	16%
anticipation-bonus	25%	38%	23%	15%

When analysing all the answers given, which are in total 2712 answers (226 participants each classifying 12 characters), four different answer patterns can be distinguished. Table 4 shows the division of answer-patterns given, split by the two different bonus-groups. As explained earlier, one group was promised a bonus for providing the most accurate classifications and the other a bonus for the most accurate anticipations. Note that from all patterns, the self-deceptive pattern (MFF or FMM) is occurring the most. Moreover, it is expected that in the anticipation bonus, there are more cases of self-deception, as this bonus-system provides a psychological incentive to answer in a self-deceptive manner. It can be seen from the table that this is indeed the case, as in the anticipation bonus group 38% of the participants answered self-deceptive, whilst in the classification bonus group this is 35%. However, with a statistic of $\chi^2 = 1.67$, the difference between these two proportions is not significant at $\boldsymbol{\alpha} = \mathbf{5}\%$.

As argued by Mijović-Prelec and Prelec, the amount of inconsistent answers (MMF or FMM) can be used as an error-baseline, in which people make an 'honest' mistake of disconfirming the first classification, and not confirming the anticipation. This can happen when somebody is 'honestly' confused during the second classification whether the character is male or female, and therefore gives two different classifications in both phases. Following the reasoning of Mijović-Prelec and Prelec (2010), the number of inconsistent patterns in this study is remarkably lower than the number of self-deceptive patterns, meaning that the self-deceptive answers cannot simply be viewed as irregularity in the classification tasks.

Before statistically analysing the data, the earlier mentioned Hausman test is performed to check whether the fixed or random effects model has to be used. The outcome of the test shows a χ^2 result of 54.73, with which the use of a fixed-effects model is confirmed.

Table 5 shows the summary statistics of the variables that are used in the regression. The dependent variable is Class2, which is the second classification. The main independent variables are Anti1, the anticipation and Class1, the first classification. R1 stands for the first confidence rating, and R2 for the second confidence rating. As stated before, Mijović-Prelec and Prelec (2010) suggest that subtracting these two confidence ratings gives an indicator of the size of the diagnostic utility. Lastly, the Group variable shows to which voting group the participants belong; the extreme or the non-extreme voters (with group 1 representing the non-extreme and group 2 the extreme). From the table it can be seen that both the dependent and the independent variables are discrete and dichotomous, as their values vary between 0 and 1 only. R1 and R2 are ordinal variables, which can take the values 1 up to 5, with 5 being the highest confidence rating, and 1 being the lowest. For all variables there are 2712 observations, meaning the dataset is strongly balanced; as for each participant (226) there are 12 observations for every variable. The mean of the Class1, Anti1 are both 0.53, meaning that 53% of the characters were classified in the first phase and anticipated in the second phase, to be male. The mean of Class2 is 0.55, meaning that 55% of the second classifications were male. From this fact solely, it can be seen that at least some of the participants changed their second classification, otherwise both means for Class1 and Class2 would be the same. Moreover, this shows that not all answers of the second classification were the same as the answer to the anticipation. The mean of the second confidence rating (3.29) is slightly higher than the first confidence rating, (3.24), meaning that the average confidence over the classification increased slightly. When performing a Wilcoxon sign test to compare these two ratings, it is found that the positive difference in confidence ratings (R2-R1), is significant (p=0.00).[1]

Table 5. Summary statistics of the variables

Variable	Obs.	Mean	Std. Dev.	Min	Max
Class1	2712	0.53	0.50	0	1
Anti1	2712	0.53	0.50	0	1
Class2	2712	0.55	0.50	0	1
R1	2712	3.24	1.04	1	5
R2	2712	3.29	1.07	1	5
Group	2712	1.11	.31	1	2

5.1 Results of the experiment

The results of the regression are summarized in table 6. This table shows the aggregate results of the logistic regression with and without adding the Group-variable.

The data of one participant was dropped, as for this participant the outcomes were all positive or negative. This means that the participant answered 'male' or 'female' for all twelve characters, for the first classification, the anticipation and the second classification. A fixed-effects model, as used in this regression, aims at finding the determinants of within-subject variability, as explained by Williams (2017). When there is no variability in the answers provided, it is not possible to apply this model to the data.

[1] The Wilcoxon sign test is used instead of the paired t-test as both the samples of R1 and R2 consist of ordinal data, and both are not normally distributed

Table 6. Regression results, with and without the interaction effect of voting-group and anticipation

	Interaction effect not included		Interaction effect included	
	Log-odds (1)	Odd-ratios (2)	Log-odds (3)	Odd-ratios (4)
Anti1	1.01*** (0.09)	2.75*** (0.24)	1.04*** (0.09)	2.84*** (0.26)
Class1	-0.84*** (0.08)	0.43*** (0.04)	-0.84*** (0.08)	0.43*** (0.04)
Group*Anti			-0.25 (0.26)	0.78 (0.20)

This table shows the results for the logit regression, both with and without the 'Group'-variable, reported in the log-odds and the odds-ratios. A * indicates significance at α=1%, **indicates significance at α=5% and *** indicates significance at α=10%.

In columns 1 and 3, the 'original' coefficients are shown, which are the coefficients without being transformed. These coefficients represent the rates of change in the 'log odds' of the dependent variable as the independent variables changes. As these coefficients are relatively difficult to interpret, taking the odds-ratio of the coefficients, which can be depicted as $exp(\beta)$, can help to understand the results better. This exponential transformation of the independent variables can be explained as the effect of the independent variable on the 'odds-ratio', the probability of the event (Class2=1) divided by the probability of the non-event (Class2=0); $(\frac{p}{(1-p)})$. The odd-ratio of the coefficients are shown in the right columns (2 and 4) of table 1. The interpretation of these ratios depends on the coding of the variables, and in this study the female characters are coded as '0', whilst the male characters are coded as '1'. The odd's-ratios measure the effect on the likelihood of answering the second classification with male, when the other two variables are answered with male too. Therefore, an odds-ratio greater than 1 describes a positive relationship, whilst an odds-ratio less than one describes a negative relationship. A negative relationship can be explained as when the answer to the independent variable (anticipation or first classification) increases (from 0 to 1), so from female to male, the likelihood that the second classification is 1 (male) decreases.

Looking at columns 1 and 2, which show the results of the regression without the interaction effect of the voting-group, it can be seen that both the first classification and anticipation are significant. More specifically, from column 2 it can be seen that without the 'Group interaction-variable', if a person switches the answer for anticipation from female to male (from 0 to 1), the odds of answering the second classification with male are multiplied by 2.75. The positive relationship, as depicted by an odd's-ratio higher than 1, indicates a positive relationship of answering anticipation with 1 and the likelihood to answer the second classification with '1' too. Additionally, if a person switches the answer for the first classification from female to male (from 0 to 1), the odds of answering the second classification with male (Class2=1) are decreased with 0.57 (1-0.43). This is explained by the odd's-ratio less than one; which depicts a negative relationship between answering the first classification with 1 and answering the second classification with 1 too.

Columns 3 and 4 show the ratios of the coefficients of the variables when the effect of voting has been added to the regression. Firstly, the odds ratios of the two independent variables Anti1 and Class1 don't change much compared to the model without the Group-variable. As can be seen from the table in the fourth column, the odds ratio from Anti1 is now 2.84, meaning that the odds of answering 'male' in the second classification are multiplied by 2.84 when the anticipation is answered with male instead of female. The Class1 odds ratio is again 0.43, meaning that answering the first classification as male reduces the odds of answering the second classification as male with 0.57.

The interaction variable measures the extent to which the group-variable influences the Anti1 effect on the second classification (Class2), adjusted for the first classification (Class1). The 'normal' coefficient can be interpreted as the ratio of the log odds, which is again not very intuitive. Taking the odds-ratio however, could give useful information for answering the hypothesis. Interpreting the coefficient of the odds-ratio of the interaction term can be formulated as the difference in odds-ratios of group 2 (the extreme voters) compared to group 1 (the non-extreme voters). When this would be a coefficient higher than 1 and significant, it would mean that the odds of extreme voters of answering the second classification with male, when the anticipation has been answered with male, are higher than those of non-extreme voters. Put differently, this interaction effect measures the difference to what extent the anticipation answer has influence on the second classification, between extreme voters and non-extreme voters. When looking at the odds-ratio interaction effect of voting-group and anticipation, a non-significant effect of 0.78 can be found. If this were to be significant, it would mean that the odds of extreme voters to influence their second classification answer by their anticipation answer are lower than non-extreme voters. Yet, the coefficient is found not to be significant, meaning that the ratio of the odds-ratios doesn't carry significant information about the tested hypothesis.

Table 7. Diagnostic utility per group-types

	Bonus-A	Bonus-C	Extreme	Non-extreme
R2-R1	0.02	0.10	0.07	0.05
n	119	107	25	201

This table shows the diagnostic utility, computed as the second minus the first confidence rating, per bonus- and voting-group. The 'n' depicts the number of participants in each group, with the total 'n' being 226.

Table 7 shows the average difference between the second and the first confidence rating, sorted by groups. According to Mijović-Prelec and Prelec (2010), the difference between those two confidence ratings is an appropriate measure of diagnostic utility, as it removes the intrinsic confidence that people have with respect to classifying the characters, as well as differences in how people use the rating scale. In the first two columns, the average diagnostic utility for the anticipation and the classification bonus group are shown consecutively. It can be seen that the average diagnostic utility for the classification-bonus group is higher than the average diagnostic utility for the anticipation bonus. As seen before, both groups have a similar number of participants. Moreover, from the last two columns it can be seen that the average diagnostic utility for extreme voters is slightly higher than for non-extreme voters, however this difference is relatively small. When performing a Mann-Whitney U test to compare those differences in confidence ratings between the two voting groups, it can be found that the difference is not significant at any level smaller than 68[2]% (p=0.68).

Table 8. Diagnostic utility per answer-pattern

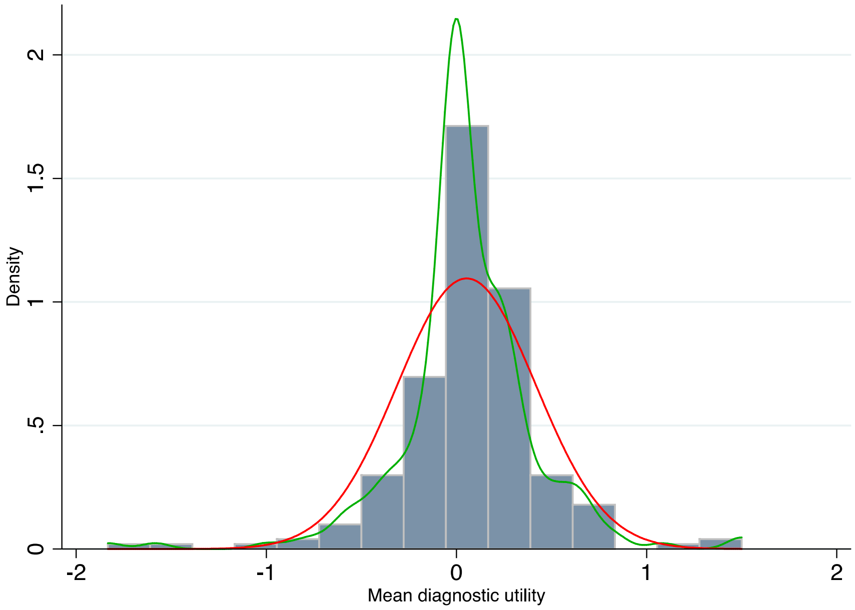
	Consistent	Self-deceptive	Inconsistent	Honest
R2-R1	0.08	-0.37	0.02	0.06
n	690	987	613	422

This table shows the diagnostic utility, computed as the second minus the first confidence rating, per answer-pattern. The 'n' depicts the number of answers in each group, with the total 'n' being 2712.

Table 8 shows the average diagnostic utilities for the different answer patterns that are given. Note that the amount of observations 'n' now depicts the amount of answers given, instead of the number of participants. In total, there are 226 participants all giving 12 answer-patterns, totaling to an amount of 2712 answer patterns. Remarkably, the average diagnostic utility gained out of the self-deceptive answers is negative, whilst all other answers are positive. This means that on average, when the second classification changes in confirmation with the anticipation answer that is given, to an answer that is not the same as the first classification, the participants are less confident about their answer. All of the other average diagnostic utilizes are positive, with the highest average being that of a consistent answer. Note that these numbers are only averages, which carry no information about the statistical significance of the differences in average diagnostic utility. Figure 2 shows the density distribution of the individual-levels of diagnostic utility. This means that the data is gathered per participant, and the mean diagnostic utility over the twelve answers is calculated. It can be seen from this figure that the distribution is quite symmetric and has long tails both on the left and the right side. The tails on the right and the left show there are some 'extreme' cases or outliers, which reflect some participants with either a 'high' or a 'low' diagnostic utility as compared to the rest. Moreover, the red line reflects a normal distribution, whilst the green line reflects the 'kernel density estimation line' that smoothens the distribution of the histogram. It can be seen that the distribution of the mean diagnostic utility is not normally distributed, as this distribution is more 'pointy-headed' and 'flat tailed'. Moreover, it can be seen that on average most participants experienced a slight positive diagnostic utility, as the peak of kernel estimation line lies just above 0 (at 0.053 precisely).

[2] A Mann Whitney U test is performed instead of a two-sample t-test as both samples (the extreme and non-extreme voters) are not normally distributed and the variable is on an ordinal scale

Figure 2. Distribution of the mean diagnostic utility



5.2 Discussion

Compared to the findings of the original experiment conducted by Mijović-Prelec and Prelec (2010), a lot more self-deceptive answer patterns are found in this research. Out of all patterns, Mijović-Prelec and Prelec found a percentage of 23.4% in the anticipation-bonus group, and an 18.3% in the classification-bonus group to be self-deceptive, whilst in this research those percentages mount up to 38% and 35% respectively. Not only is the size of these numbers different, Mijović-Prelec and Prelec also found that participants in the anticipation-bonus group answered with self-deceptive answers 5.1% more than in the classification-bonus group. In our research this difference is at a non-significant 3%.

When looking at the methodologies, some differences between the two experiments can be indicated. Whilst Mijović-Prelec and Prelec conducted the experiment in real life, adding some kind of 'interviewer effect' (which can be compared to the face-to-face interview-effect of polling methods), this research is completely conducted online. As stated in the literature review, with face-to-face interview methods people are more prone to be subject to the 'social desirability bias', causing them to give more social desirable answers. It could be the case that a comparable effect is also present in answering this experiment. In this way the answer patterns can be answered in a 'desirable way', which would be recognized as the consistent and honest patterns, which consist of the same first and second classification (meaning the participant doesn't disagree with him-/herself). In this way, those two patterns can be seen as 'true' patterns, especially when it's assumed that the participants are able to recall their first classification in the second part of the experiment. The experiment in this research is conducted online and from home, meaning that the participants are not subject to this so-called social desirability bias. This could one of the reasons for the high percentage of self-deceptive answer patterns.

Moreover, the reward structure of this research is different from the one used in the original experiment. In the original experiment, the participants got rewarded with actual money for each correct anticipation and classification they provided, so the chances of winning at least something were high. In this experiment however, participants are rewarded with points, and only the 5 people with the highest points will actually win a prize. In this matter, the chances of winning a prize feel 'further away' than in the actual experiment. It could be that also for this reason, participants are more incentivized to be self-deceptive, as the chance of winning something is much lower than in the original experiment. Holt and Laury (2002), investigated the risk aversion in an experiment using a Random Lottery Incentive, in which randomly one task or one participant is chosen to be actually paid. They found that if there are too many tasks (or too many participants), there is a low probability that a specific task or participant is selected, which can cause the subjects to not take the tasks seriously. Moreover, as found by Camerer et al. (1999), higher levels of incentives have the largest effects in judgment en decision tasks. Following this line of thought, as this experiment is a type of judgment task, relatively high and realistic incentives should be used.

Looking at the results of the model with the interaction effect included, both the Anti1 and the Class1 variables are found to be significant. The odds-ratio for the anticipation sign is 2.84, meaning that the odds of answering the second classification with 'male' when the anticipation was answered with male are multiplied by 2.84. Put differently, the answer to the anticipation increases the odds of providing a confirming second classification with 184%, regardless of whether this second classification is in line with the first classification, ceteris paribus. This means that on aggregate level, looking at all the participants at once, the 'anticipation' is significant, which is the first sign of self-deceptive participants in the dataset. This result is in line with the earlier findings that showed the self-deceptive answer pattern to be the most popular pattern in this dataset. Mijović-Prelec and Prelec conduct a different analysis, as they run individual logistic regression to determine on an individual basis whether a participant is considered self-deceptive or not (by looking if Anti1 is significant or not). Therefore, they do not state the exact coefficient of Anti1, as this is not the focus of their study.

The odds-ratio for the first classification (Class1) is 0.43, which can be interpreted as the answer to the first classification reducing the odds of providing a confirming second classification with 57%, ceteris paribus. As the tables provide the results on aggregate level, this 'low' odds-ratio can be caused by the fact that some participants don't recall their first classification answer, and answer randomly when they are shown the same characters for the second classification. However, it has to be noted that in this research the participants are only shown twelve characters, which is a big reduction compared to the original experiment of Mijović-Prelec and Prelec (2010) in which the participants saw 100 characters. This result can also be interpreted as the fact that participants are more influenced by their answer to 'anticipation' than their answer to the first classification. This can be due to the same reasons as to why the amount of self-deceptive answers in this research is high compared to the original experiment.

The odds-ratio for the variable of interest, the interaction term between voting-group and anticipation, is found to be insignificant and a level of 0.78. This means that from this research, there is not enough evidence to reject the null-hypothesis, that extreme voters are equally self-deceptive as non-extreme party voters. This is confirmed when looking at the size of the average diagnostic utilities for the two groups, which are 0.07 and 0.05 for extreme and non-extreme voters consecutively. If extreme voters were more self-deceptive, it would also be expected that their diagnostic utility would be higher, as it is the diagnostic motivation that drives self-deception. However, the difference in diagnostic utility between the extreme and non-extreme voters is small. However, the measure of diagnostic utility is generated from 12 observations only, compared to 100 in the case of Mijović-Prelec and Prelec. It could be the case that this measure is therefore biased, and further research should be conducted in order to make valid conclusions.

Another interesting finding is the negative average diagnostic utility that is found when participants provide a self-deceptive answer pattern. The negative average utility that is found can be explained as the fact that after providing a self-deceptive answer, participants on average state a lower second confidence rating than their first confidence rating. This is also in line with the finding that in the anticipation-bonus group, where there are more cases of self-deception, the overall diagnostic utility is lower than in the classification-bonus group (0.02 compared to 0.10).

As mentioned in the methodology, diagnostic utility can be formally depicted as $E(\theta_M|x = m) - E(\theta_M\ |x = f)$ or as $E(\theta_F|x = f) - E(\theta_F\ |x = m)$. As stated before, the diagnostic utility is the difference in expectation of the deep belief (this deep belief is either male or female) inferred from either classifying the character (for the second time) as male or female. It measures the extent to which self-deceptive behavior gives the participant some extra utility, or psychological benefit. The negative average diagnostic utility of -0.37 means that by providing a self-deceptive answer pattern, the participants experience on average not a positive or extra utility, but rather negative and thus decreasing utility. This could be the case as the participant is aware of the fact that he/she is not answering in line with the deep belief. However, it has to be kept in mind that only the self-deceptive answer pattern causes a negative utility on average. It could still be the case that on an individual level, the participant experiences an overall positive diagnostic utility. This way of analyzing the results is different than the method of Mijović-Prelec and Prelec, who conducted an individual-level regression, and observed for each participant whether the anticipation-coefficient was significant; for which person would be classified as self-deceptive. As in this research only 12 characters are tested instead of 100, the sample for an individual-level regression per participant is too small, so only the aggregate results are provided. When looking at the distribution of the individual average diagnostic utilities, it is found that most participants experience a small but positive diagnostic utility. Although their self-deceptive answers cause on average a negative diagnostic utility, their total diagnostic utility stays above 0.

6. Conclusion

The discrepancy between political polling predictions and the actual outcome of elections has been a much-discussed topic worldwide. How can it be that with technological advances which make it easier to conduct polls a vast number of polling predictions are still far from accurate? With the increased use of mixed-method polling techniques that rule out the bias of each separate technique, the reason for the wrong predictions is assumed to be related to the provided answers rather than the poll methods itself. More specifically, in this paper the psychological practice of 'self-deception' is used as an alternative explanation for the discrepancy between poll predictions and outcomes. By using self-deception as an alternative explanation, it is argued that people (aware or unaware) misreport their voting intentions in order to send a positive signal to themselves. As extreme left or right-wing parties are often portrayed in the news as being controversial, it could be the case that although someone's deep belief is in line with that such an extreme party, the person chooses to report a different voting intention, as he/she does not want to confirm this deep belief that moment in time. However, on the election day the deep belief wins, and the person votes for the party he/she truly supports. This research aimed at finding out whether voters for extreme parties, either extreme right or left, are more prone to show self-deceptive behavior than non-extreme party voters. An experiment was conducted in order to find evidence for the relation between self-deception and extreme voting preferences. In total, 226 people participated in the actual experiment, who were gathered through various channels like Twitter and Facebook, trying to reach as many and as diverse people as possible. The experiment was for a great extent based on the experiment of Mijović-Prelec and Prelec (2010). However, the experiment of this paper was conducted online instead of in real life and involved substantially lower incentives for the participants. Based on the outcomes of the experiment, there was not enough evidence found to conclude that extreme party voters are more likely to show self-deceptive behavior. However, the outcomes do show a high amount of self-deceptive answers, meaning that self-deception was present in the answers provided. More specifically, in both bonus conditions (the anticipation and the classification bonus) the self-deceptive answer pattern was the one that occurred the most (38% and 35% respectively). As the incentive to provide such a self-deceptive answer was purely psychological, in the case of the classification-bonus more so, this does raise the question why people to a large extent show such behavior. Is it because the incentive system of the experiment was unclear to them, and they truly believed that a self-deceptive answer pattern would mean more points and thus more change of winning a voucher? Or is it the case that confirming one's anticipation and ignoring the deep belief gives a psychological spur that one is doing a good job? The high amount of self-deceptive answers combined with the finding that on average this answer pattern causes a negative diagnostic utility (is the participant aware of being 'untrue' and therefor experiences a negative utility?) raises interesting questions that could be explored in future research.

6.1 Limitations and recommendations

The methodology and data used in this experiment have some limitations that could alter the outcomes and effect sizes. Firstly, the number of participants that voted for 'extreme' parties, 25, is very low compared to the non-extreme parties, 201. Also, when looking at the actual votes per party in Dutch politics, the sample in this experiment does not correctly reflect that dispersion. One reason for this limitation could be that the experiment was distributed online only, reaching a limited amount of people. It would be advisable for future research to conduct the experiments both online and in real life, for example on the streets. Including a higher number of extreme party voters would be recommended as well, as a higher sample size leads to more accurate statistical results. Moreover, as mentioned before, the reward that was gained in this experiment was quite small. In the real experiment by Mijović-Prelec and Prelec, all participants were rewarded in any case, whilst in this experiment only 5 participants were rewarded a voucher. This could lead to the fact that the participants take the experiment less serious. It is advisable to reward all participants with the amount of points they 'won' in the experiment, to remove any extra psychological motivation that is caused by a reward that is too unreachable. Related to this limitation is the number of characters shown to the participants. In this experiment, only 12 characters are shown, in comparison to 100 characters in the original experiment. The reason for limiting the amount of characters is linked to the low rewards, which makes it hard to find participants who are willing to participate in an experiment which takes long. Hence, with this small sample size consisting of only 12 observations per participant, it is unreliable and statistically incorrect to conduct an analysis per participant. For this reason, only an aggregated analysis is possible. It would be interesting for future research to include more characters, and to perform an individual level analysis as well. This could then also make it easier to include statistics that correct for the serial correlation that might be included in the estimators of this experiment. Lastly, by using only twelve characters in the experiment, there is more chance that outcome utility drives the motivation (measured by the difference in confidence ratings) as well. This is the case as the participants are more likely to remember the twelve characters (as compared to 100) from phase 1, and therefore can predict with a higher probability than 0.5 whether some characters in phase 2 will be female or male. To rule out outcome utility and to only facilitate diagnostic utility, the number of characters must be high enough for the participants to remember that they answered (and how often they answered male/female) in phase 1. Future research could implement a sort of time-limit per answer that prevents participants for remembering their answers to phase 1.

References

Bem, D. J. (1972). Self-perception theory. *Advances in experimental social psychology*, *6*, 1-62.

Berinsky, A. J. (2006). American Public Opinion in the 1930s and 1940s The Analysis of Quota-Controlled Sample Survey Data. Public Opinion Quarterly, 70(4), 499-529.

Blumberg, S. J., & Luke, J. V. (2007). Coverage bias in traditional telephone surveys of low-income and young adults. Public Opinion Quarterly, 71(5), 734-749.

Blumenthal, M. M. (2005). Toward an Open-Source Methodology What We Can Learn from the Blogosphere. Public Opinion Quarterly, 69(5), 655-669.

Bodner, R., & Prelec, D. (1997). The diagnostic value of actions in a self-signaling model, MIT mimeo, January.

Bodner, R., & Prelec, D. (2003). Self-signaling and diagnostic utility in everyday decision mking. The psychology of economic decisions, 1, 105-26.

Burke, J., & Taylor, C. R. (2008). What's in a poll? Incentives for truthful reporting in pre-election opinion surveys. Public Choice, 137(1), 221-244.

Camerer, C. F., Hogarth, R. M., Budescu, D. V., & Eckel, C. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. In Elicitation of Preferences (pp. 7-48). Springer Netherlands.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. The Public Opinion Quarterly, 64(4), 464-494.

Crespi, I. (1988). Pre-election polling: Sources of accuracy and error. Russell Sage Foundation.

De Leeuw, D. (2005). To mix or not to mix data collection modes in surveys. Journal of official statistics, 21(2), 233.

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. Journal of consumer research, 20(2), 303-315.

Fuchs, M., & Busse, B. (2009). The coverage bias of mobile web surveys across European countries. International Journal of Internet Science, 4(1), 21-33.

Gallup, G. (1957). The changing climate for public opinion research. Public Opinion Quarterly, 21(1), 23-27.

Goeree, J. K., & Grosser, J. (2007). Welfare reducing polls. Economic Theory, 31(1), 51-68.

Gur, R. C., & Sackeim, H. A. (1979). Self-deception: A concept in search of a phenomenon. Journal of Personality and Social Psychology, 37(2), 147.

Henderson, B. (2016, November 8). Hillary Clinton or Donald Trump? America starts to vote in the most divisive election in history: Tuesday US election briefing. Telegraph. Retrieved from: http://www.telegraph.co.uk/news

Hillygus, D. S. (2011). The evolution of election polling in the United States. Public opinion quarterly, 75(5), 962-981.

Hogan, J. M. (1997). George Gallup and the rhetoric of scientific democracy. Communications Monographs, 64(2), 161-179.

Holt, C. A., & Laury, S. (2002). Risk aversion and incentive effects. The American Economic Review, 92 (5), pp. 1644-1655

Jowell, R., Hedges, B., Lynn, P., Farrant, G., & Heath, A. (1993). The 1992 British election: the failure of the polls. The Public Opinion Quarterly, 57(2), 238-263.

Keeter, S. (2006). The impact of cell phone noncoverage bias on polling in the 2004 presidential election. Public Opinion Quarterly, 70(1), 88-98.

Marks, G., Wilson, C. J., & Ray, L. (2002). National political parties and European integration. American Journal of Political Science, 585-594.

Mijović-Prelec, D., & Prelec, D. (2010). Self-deception as self-signalling: a model and experimental evidence. Philosophical Transactions of the Royal Society B: Biological Sciences, 365(1538), 227-240.

Mokrzycki, M., Keeter, S., & Kennedy, C. (2009). Cell-phone-only voters in the 2008 exit poll and implications for future noncoverage bias. Public Opinion Quarterly, 73(5), 845-865.

Pickup, M., Matthews, J. S., Jennings, W., Ford, R., & Fisher, S. D. (2011). Why did the polls overestimate Liberal Democrat support? Sources of polling error in the 2010 British general election. Journal of Elections, Public Opinion and Parties, 21(2), 179-209.

Pindyck, R. S., & Rubinfeld, D. L. (1998). Econometric models and economic forecasts (Vol. 4). Boston: Irwin/McGraw-Hill.

Sears, D. O., & Funk, C. L. (1990). The limited effect of economic self-interest on the political attitudes of the mass public. Journal of Behavioral Economics, 19(3), 247-271.

Shepard, (2015, July 10). Gallup gives up the horserace. Politico. Retrieved from: https://www.politico.com/story/2015/10/gallup-poll-2016-pollsters-214493

Siemiatycki, J., & Campbell, S. (1984). Nonresponse bias and early versus all responders in mail and telephone surveys. American Journal of Epidemiology, 120(2), 291-301.

Stock, J. H., & Watson, M. W. (2007). Econometrics. Addison- Wesley.

Stout, C., & Kline, R. (2008). Ashamed not to vote for an African-American; ashamed to vote for a woman: An analysis of the Bradley effect from 1982-2006. Center for the Study of Democracy.

Streb, M. J., Burrell, B., Frederick, B., & Genovese, M. A. (2007). Social desirability effects and support for a female American president. Public Opinion Quarterly, 72(1), 76-89.

Struthers, J., & Young, A. (1989). Economics of voting: Theories and evidence. Journal of Economic Studies, 16(5).v

Terhanian, G. (2008). Changing times, changing modes: the future of public opinion polling?. Journal of Elections, Public Opinion and Parties, 18(4), 331-342.

Torres-Reyna, O. (2007). Panel data analysis fixed and random effects using Stata (v. 4.2). Data & Statistical Services, Priceton University.

Traugott, M. W. (2005). The accuracy of the national pre-election polls in the 2004 presidential election. Public Opinion Quarterly, 69(5), 642-654.

Trivers, R. (2000). The elements of a scientific theory of self-deception. Annals of the New York Academy of Sciences, 907(1), 114-131.

Quattrone, G. A., & Tversky, A. (1986). Self-deception and the voter's illusion. The multiple self, 35-38.

Williams, R. (2017, January 22). Serial Correlation. Retrieved from: https://www3.nd.edu/~rwilliam/stats2/l26.pdf

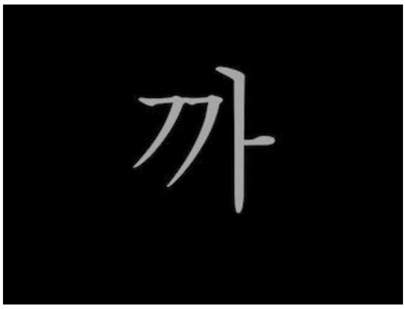
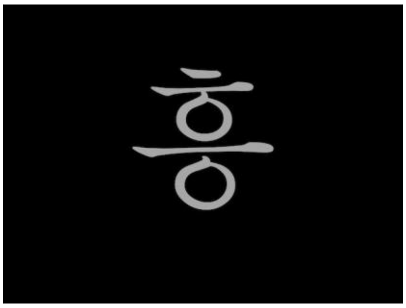
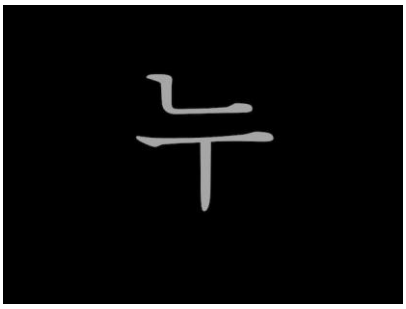
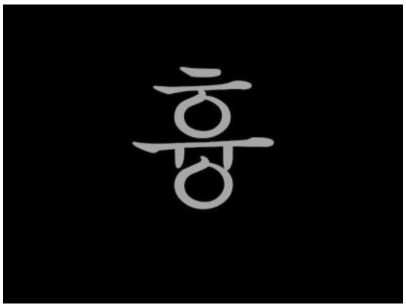
Appendix A

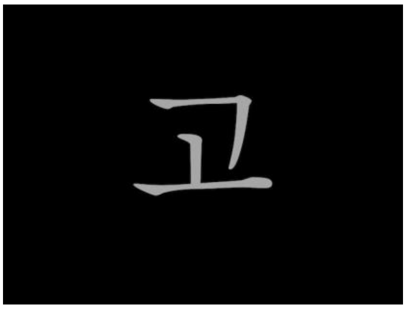
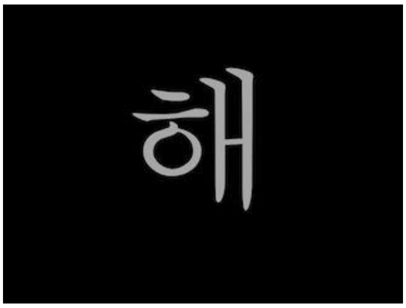
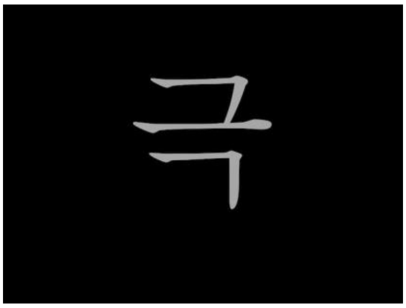
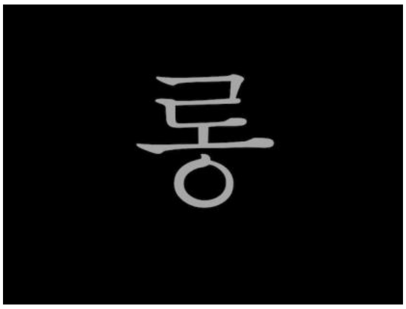
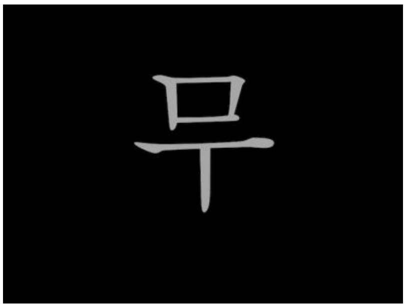
Overview of all political parties in the elections of the Netherlands 2017, the number of votes received (in numbers and percentages), and the number of seats

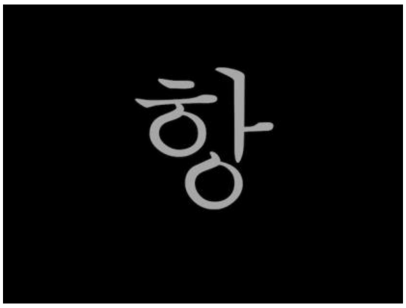
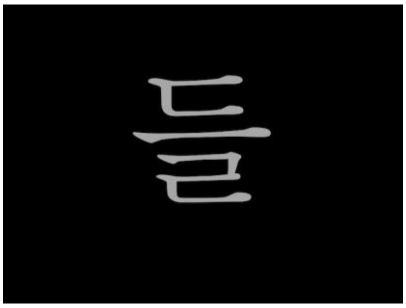
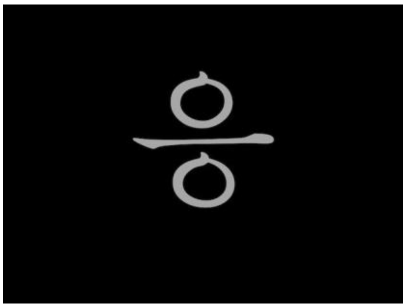
Party	Votes	Percentage	Seats
VVD	2.238.351	21,3	33
PVV	1.372.941	13,1	20
CDA	1.301.796	12,4	19
D66	1.285.819	12,2	19
GroenLinks	959.600	9,1	14
SP	955.633	9,1	14
PvdA	599.699	5,7	9
ChristenUnie	356.271	3,4	5
Partij voor de Dieren	335.214	3,2	5
50Plus	327.131	3,1	4
SGP	218.950	2,1	3
DENK	216.147	2,1	3
Forum voor Democratie	187.162	1,8	2
VNL	38.209	0,4	-
Piratenpartij	35.478	0,3	-
Artikel 1	28.700	0,3	-
Nieuwe Wegen	14.362	0,1	-
Ondernemerspartij	12.570	0,1	-
Lokaal in de Kamer	6.858	0,1	-
Niet Stemmers	6.025	0,1	-
De Burger Beweging	5.221	-	-
GeenPeil	4.945	0,0	-
Jezus Leeft	3.099	0,0	-
Vrijzinnige Partij	2.938	0,0	-
LP	1.492	0,0	-
MenS en Spirit/BP-VR	726	0,0	-
StemNL	527	0,0	-
VDP	177	0,0	-
Totaal	10.516.041	100%	150

Appendix B

The table underneath shows the twelve characters with the highest conformity rate from the pre-experiment

Character	**Conformity**
	**Male** : 81% Female : 19%
	Male : 14% **Female** : 86%
	**Male** : 84% Female : 16%
	Male : 14% **Female** : 86%

	**Male** : 87% Female : 13%
	Male : 19% **Female** : 81%
	**Male** : 84% Female : 16%
	Male : 17% **Female** : 83%
	**Male** : 88% Female : 12%

	Male : 10% **Female** : 90%
	**Male** : 84% Female : 16%
	Male : 12% **Female** : 88%