MASTER THESIS

A non-monetary truth-telling incentive

Luc Schneider – 381324
Supervisor: Aurelien Baillon
Department of Applied Economics
Erasmus School of Economics

Date: 19-07-2017

Abstract

In this paper, I compare the effect of Bayesian Truth Serum to that of an Honour code, which serves as a non-monetary truth-telling incentive, on admission rates to unverifiable and questionable behaviour. I collected data from 736 students of the Erasmus University of Rotterdam on six forms of questionable behaviour through an online survey. I find that both the BTS and the Honour code (self-concept treatment) elicit higher admission rates to questionable behaviour than the control survey. No significant differences are found between the two truth-telling mechanisms.

Table of content

1. Introduction

When asking people personal questions, there are many things that are completely unverifiable. For instance, opinions, or certain types of questionable behaviour cannot be checked. *A priori*, the only way to find out about such things is to ask the person, and to trust that he or she answers truthfully. For research purposes, however, this may not always be satisfactory.

In recent years, studies have become interested in different types of mechanisms that could stimulate higher degrees of truth-telling on verifiable or unverifiable questions. Studies in psychology (Dickerson et al., 1992; Mazar et al., 2008), or in various other fields (John et al., 2012; Weaver & Prelec, 2013) have started to compare different mechanisms to assess their relative efficiency.

From these two strands of literature, two different approaches emerged: on the one hand, economists designed monetary truth-telling incentives by scoring answers in a way that made truth-telling the utility-maximising response for Bayesian updaters (Prelec, 2004), while on the other hand, researchers in psychology tried to stimulate truth-telling by using concepts such as cognitive dissonance (Mazar et al., 2008).

These approaches were later compared to each other by Weaver & Prelec (2013) in the context of truth-telling on verifiable questions. They found that the monetary incentives clearly dominated the psychological ones. However, no such comparison was made when considering questions of a different nature – that is, unverifiable questions regarding questionable or shameful behaviour.

In the present paper, two of the treatments used by Weaver & Prelec (2013) are compared to each other again when asking participants about socially unacceptable behaviour. In doing so, the experimental design combines the treatments of Weaver & Prelec (2013) and the question format of John et al. (2012) in an attempt to draw conclusions that can be related to previous literature, thus contributing to the ongoing expansion of this strand of research.

A survey was conducted predominantly amongst students of the Erasmus University of Rotterdam. All respondents answered six questions regarding questionable behaviour,

and were assigned to either a control group or one of two treatments. Results indicate that each of these treatments yielded higher admission rates than the control group.

This paper is structured as follows: first I present a discussion of previous literature on truth-telling, self-concept, and the Bayesian Truth Serum. This is followed by a section covering the hypotheses that are tested, along with a presentation of the survey design. After this, a data section gives an overview of the basic findings resulting from the experiment, as well as a set of descriptive statistics. Then, a brief methodology discusses the statistical approach to answering the hypotheses. Finally, I present the results and draw implications regarding the effectiveness of the considered truth-telling mechanisms.

2. Literature Review

The literature review consists of an overview of the various concepts and theories that are used in the construction of the survey experiment. First, I investigate the self-concept literature in order to introduce the theory of self-concept maintenance. Following this, I address self-deception, and I give an overview of the various ways in which self-concept maintenance can be used to stimulate truth-telling. Finally, I discuss truth-telling mechanisms directly, with special focus on the Honour code mechanism and on the Bayesian Truth Serum.

2.1 Self-concept

Self-concept, along with its role in decision-making and behaviour, has been extensively discussed in various strands of literature, including marketing, economics, sociology and psychology. In the following parts, I give an overview of this literature to gain a broader understanding of how self-concept influences the way people act. First, I discuss some background, and a formal definition of the self-concept. Subsequently, I present the characteristics of the self-concept as well as the theory of self-concept maintenance. Finally, I take a closer look at the role of self-deception in self-concept maintenance.

2.1.1 Background and definitions

Self-concept can be broadly defined as “the way people view and perceive themselves” (Aronson, 1969; Bem, 1972; Mazar et al., 2008). The discussion around self-concept, before being an academic one, was predominantly philosophical. Sartre (1956)

extensively discussed the matter of self. He argued that the self cannot be evaluated "unless it is the object of thought". This assessment later led Ickes et al. (1973) to distinguish between objective and subjective self-awareness. In this case, objective self-awareness describes an introspective state in which the individual is capable of self-evaluation. Subjective self-awareness, on the other hand, is a state that does not allow for self-evaluation.

The authors recognised two key characteristics of these states of self-awareness: (1) an individual cannot be in both states at the same time, meaning that he is, at any point in time, either objectively or subjectively self-aware, but never both; and (2) objective self-awareness can be stimulated, which also encourages self-evaluation (Ickes et al., 1973).

They argued that self-awareness is used to update the self-concept, and that only the objectively self-aware individual will notice any differences between his (moral) standards, or goals, and the implications of his present condition. A subjectively self-aware person will not see these differences, and will thus not need to correct for this discrepancy between his actions and his beliefs (Ickes et al., 1973).

However, the second characteristic of these self-awareness states points towards the possibility to influence someone's propensity to self-evaluate, indicating that the updating of the self-concept can by stimulated (Ickes et al., 1973). A number of experiments (Gergen, 1965; Walster, 1965, 1970) seem to confirm this. These experiments showed that direct feedback can be used to induce changes in self-concept when the individual is confronted with personal or social norms.

2.1.2 Social and personal norms

The self-concept, as defined in the previous part, depends heavily on personal norms. Indeed, the way people perceive themselves depends on the standards of behaviour that people adopt. This allows people to form their self-concept based on the comparison between their actions and this standard of behaviour. In this area, psychologists have argued that while personal norms guide behaviour, social norms play an important role in determining personal norms (Campbell, 1964).

More specifically, it would appear that people naturally internalise social norms and values, and that they subsequently use them to form their personal norms. Self-concept

and behaviour are thus driven both by social and personal norms, in that the latter is directly influenced by the former. This idea was applied in an experiment by Henrich et al. (2001), who gathered data on traditional economic games in fifteen "isolated" societies. The experiment confirmed that personal behaviour varies greatly between societies, which confirmed the idea that social norms play an important role in determining behaviour.

Furthermore, the dependence of the self-concept on both social and personal norms indicates that people's self-concepts are generally influenced by certain common moral standards, such as honesty. As such, if a given society strongly values honesty, it can be expected that virtually all individuals in this society strongly value honesty as well. This has been confirmed by various studies, along with the fact that most people consider themselves to be honest (Mazar et al., 2008, Beznea et al., 2016).

These norms and values are then assessed and used by an internal reward system to determine how an individual feels about a certain action.

2.1.3 Internal reward system

Mazar et al. (2008) went on to discuss how this behaviour that is influenced by social and personal norms affects the individual by stimulating his internal reward mechanism. Essentially, they argued that the internal reward system that activates when an individual receives an external reward, such as money, or candy, is also activated when the individual acts according to his norms and values. Similarly, when the individual violates these norms, this will be perceived in similar ways to a punishment.

De Quervain et al. (2004) and Rilling et al. (2002) found evidence of this in neuroscience, as they specifically looked at how the impact of social norms on behaviour activates certain parts of the brain. They found that brain activity around upholding or violating social norms is associated with the same part of the brain that processes extrinsic rewards, thus confirming that the comparison between behaviour and norms, and therefore, self-concept, is associated with the individual's internal reward system.

These findings highlight the importance of the discussion around self-concept, and, more specifically, around how awareness of the self-concept influences the individual. If self-

concept is associated with the internal reward system, understanding how the self-concept is influenced by behaviour is essential.

2.1.4 Self-concept maintenance

The previously discussed literature indicates that when an individual is objectively self-aware, an action violating his self-concept will deplete it, while an action upholding the norms and values that are central to this self-concept will comfort or even enhance it. In line with this, Mazar et al. (2008) formulated a theory of self-concept maintenance. The theory suggests that people tend to be dishonest and lie only to the extent that they can justify this without compromising the image that they have of themselves (i.e. self-concept).

This theory has a number of important implications. First, it implies that people might be more inclined to be dishonest or to lie when they are subjectively self-aware, as this state allows for larger discrepancies between actions and beliefs than when people are in a state of self-evaluation (Ickes et al., 1973). Therefore, if the state of self-awareness influences the self-concept, and if the self-concept, in turn, is linked to the internal reward system, the ability to influence the state of self-awareness should also affect the individual's behaviour, through his self-concept.

Second, it means that people's behaviour also depends heavily on the way the action is perceived. One can think of a large number of actions that are unambiguously acceptable or unambiguously bad, but there also exists a large range of actions or behaviours that are not as easily classified, and that vary widely in terms of malleability (Mazar et al., 2008). When it comes to these types of actions, the literature finds that higher degrees of malleability make it easier to justify lies or dishonest behaviour.

This is confirmed by Cunha and Cabral-Cardoso (2006), who recognised that people face this malleability dilemma even when it comes to formal laws and rules, which should technically be stronger than mere norms of behaviour. Indeed, they argued that situations may arise in which not following the rules is actually the better thing to do. The choice then becomes a matter of deciding whether the rule supersedes the need for breaking it. This point is especially interesting as it shows the limits of the rules people are confronted with. Just like legal rules are subject to these limits, social and personal norms may be subject to them too. These limits may justify violating some of these norms.

It thus becomes clear that there is some sort of "grey area", in which the range of behaviours is of unclear nature; and the ambiguity of such actions then means that it is easier for the individual to maintain his self-concept while still acting against social or personal norms (Mazar et al., 2008).

Furthermore, the way a certain action is perceived can differ depending on whether the individual considers the medium or the outcome resulting from this action. Hsee et al. (2003) argued that most effort typically leads to the acquisition of a medium, which is then used to reach the desired outcome. However, it appears that people tend to overvalue the medium compared to the outcome. This phenomenon is known as psychological myopia, and may cause people to disguise an outcome that would hurt their self-concept by focusing on the medium (Hsee et al., 2003). Such behaviour may, to a certain extent, be categorised as self-deception, which plays a key role in self-concept maintenance.

2.1.5 Self-deception

Sartre (1956) said that the capacity to self-deceive oneself necessarily implies two contradictory things: (1) that an individual can hold beliefs that he is not aware of; and (2) that the individual must know all of his beliefs sufficiently well to understand how to hide these beliefs from himself, thus deceiving himself. This paradox evokes some interesting thoughts regarding how self-deception works, but also regarding how self-deception influences self-concept maintenance.

Following Sartre (1956), Gur and Sackeim (1979) argued that self-deception could be categorised into four steps. First, the individual must hold two beliefs that are mutually exclusive, such that believing one means that he cannot believe the other. These two beliefs must be held simultaneously, and the individual must be unaware of one of them, thus satisfying Sartre's first interjection. Finally, "the non-awareness of this belief must be motivated", which approximately satisfies Sartre's second implication. The authors proceeded to test these steps using an experiment that measured both conscious and unconscious beliefs by using sensory and regular reaction to two types of voice recordings (Gur & Sackeim, 1979).

The authors exposed subjects to recordings of their own voice, and to the voice of strangers, and asked these subjects to identify who the voice was from, while applying

sensors to measure the body's reaction to the recordings. In this setting, the sensor data provided evidence of beliefs that the subject is unaware of, while the direct answers provided evidence of the conflicting belief when the subject was wrong in identifying his own voice or the voice of someone else. Although the evidence for the fourth step of the self-deception process, which concerned the motivation of the non-awareness, was weak, the results supported all three others (Gur & Sackeim, 1979). This confirmed the premise that people can hold two different "beliefs" while not being aware of one of them.

While not directly related to the self-concept, these findings provided crucial insights in the ability of people to deceive themselves. Applying these findings to self-concept maintenance, it confirms that there is room for people to remain unaware of certain aspects of their self-concept, allowing for discrepancies between behaviour and norms without jeopardising the maintenance of the self-concept. This would be especially true when individuals are not objectively self-aware, as described by Ickes et al. (1973). Research on cognitive dissonance, which describes a state in which an individual holds two contradicting beliefs at the same time (Aronson, 1969), further confirmed this.

Self-deception, as characterised by the steps of Gur and Sackeim (1979), can take on various forms. For instance, one of the ways in which people self-deceive is by distorting reality. Griffin and Ross (1991) recognised that people perceive reality subjectively, creating personal representations of events and actions around them. This explains the focus on both social and personal norms in the composition of the self-concept; instead of merely social norms. For instance, it allows people to differ in what they consider to be honest or dishonest. While it is clear that people value honesty, the subjectivity of human perception allows people that might objectively be considered dishonest to maintain an honest self-concept (Griffin & Ross, 1991; Greenwald, 1980).

Furthermore, Greenwald (1980) introduced the concept of beneffectance, which states that people are predisposed "to take credit for desirable outcomes while denying responsibility for undesirable ones". As such, not only do people judge reality subjectively, they will favour subjective assessments that make themselves look better, both to themselves and to others.

On top of this, self-deception can be stimulated. When people are led to believe that a certain trait is particularly desirable, they will naturally look for this trait in their

personal memories. These memories will then be used to justify the desired self-concept, showing that this self-concept depends on people's subjective assessment of memories (Sanitioso et al., 1990). In particular, the study by Sanitioso et al. (1990) showed that subjects were more likely to associate themselves with the characteristics that the researchers arbitrarily chose to make desirable. They concluded that people are constrained by the content of their memories to create the self-concept that they desire, but that the subjective nature of these memories makes self-deception possible.

2.2 Using self-concept maintenance to stimulate truth-telling

After discussing the self-concepts and the various ways it is used and influenced by internal and external factors, I focus on how to use the self-concept maintenance theory to stimulate truth-telling. First, I give an overview of the literature concerning lie aversion to assess the extent to which people are dishonest in general. Next, I present commitment devices as a way to reconcile actions and self-concept.

2.2.1 Lie aversion

Standard economic theory predicts that people will lie and be dishonest as long as this is more beneficial than telling the truth. From a utility standpoint, this seems to make sense. However, as established previously, honesty is both a social and a personal norm, and several studies (Gneezy, 2005; Lundquist et al, 2009; Maas & Van Rinsum, 2012) suggested that lying or deceiving others decreases an individual's utility.

In particular, Gibson et al. (2005) and Lundquist et al. (2009) found that, even in settings where there is no social or economic downside to lying, some subjects choose not to lie, and those that lie rarely do so to the maximum extent. Even beyond this, findings suggest that people who are presented with an economic incentive to lie often still favour small lies over big ones, even though big lies would clearly make them economically better off (Lundquist et al., 2009).

This suggests that lying comes at a clear psychological cost, as it indicates that people are willing to forego economic gains to avoid lying. It also confirms that dishonesty forces the individual to update his self-concept, at least to the extent that he is aware of it (Mazar et al., 2008). Furthermore, it establishes that there are limits to the effectiveness of self-deception in maintaining the self-concept, as this is also affected by the magnitude of the lie (Schweitzer & Hsee, 2002).

Finally, the context in which people lie clearly affects their disposition to do so as well. As Maas and Van Rinsum (2013) exposed, an individual will be less likely to lie, and will lie to a smaller extent, when he is aware that the lie will be shared in a social context. Lie aversion can thus vary depending on the social context. In fact, Mazar et al. (2008) found that people tend to lie about their behaviour only when they operate under the belief that this behaviour is completely unverifiable, and when they are not under the influence of any self-concept deterring mechanisms.

2.2.2 Cognitive dissonance and commitment as self-concept deterring mechanisms

The theory of cognitive dissonance predicts that, when people hold two contradicting beliefs about their behaviour, they will attempt to bridge the gap between those two beliefs by choosing to behave in the way that is closest to their norms and values (Aronson, 1969). Cognitive dissonance typically arises when an individual is reminded of the discrepancy between his behaviour and the "right" behaviour. This makes him aware of how his actions hurt his self-concept, and as a result, he has to either correct his actions or update his self-concept. Since the latter is psychologically costly, cognitive dissonance is expected to induce behavioural change (Aronson, 1969).

Aronson (1980) further argued that when cognitive dissonance involves someone's self-concept, the individual must rethink or justify his actions to himself, which is most easily done by adopting behavioural changes. In other words: making someone aware of their self-concept may be an efficient way of inducing changes in behaviour, both because of the cost of updating one's self-concept and because of the cognitive dissonance mechanism.

The effects of this mechanism were tested in various environments. For instance, Cohen et al. (1963) found evidence of behavioural changes due to cognitive dissonance in social environments. Although these changes are only measured in the short-run, Freedman (1965) also found long-term effects of cognitive dissonance on the behaviour of small children. In general, the literature finds cognitive dissonance to be an effective motivator to push people to behave in the (socially) desired way (Dickerson et al., 1992).

One specific application of cognitive dissonance uses commitment devices to highlight the discrepancy between people's actions and their beliefs. The intuition behind this is

that the discrepancy becomes more evident when people not only are made aware of it, but also agree to act in accordance to the desired behaviour. According to Dickerson et al. (1992), this creates a powerful feeling of hypocrisy within subjects that previously did not behave the "right" way. This feeling of hypocrisy reinforces the need to compensate for the wrong behaviour to maintain the self-concept.

This type of commitment device was found to be highly effective in reducing dishonesty as well. Mazar et al. (2008) found that making subjects sign an honour code virtually eliminated dishonesty and lying about performance. Furthermore, research on academic integrity confirmed that honour codes, when they serve as commitment devices, lead to significantly less academic dishonesty (McCabe & Trevino, 1993; McCabe et al., 2002). Bok (1990) even stated that honour codes may be the best way to stimulate academic integrity, both among students and staff.

Mazar et al. (2008) argued that the honour code serves as a reminder of morality, which affects the respondent's ability to maintain his self-concept. By using such a reminder of morality, they increased the standard of honesty of the subjects, which deterred them from lying by arousing cognitive dissonance. Across a number of other experiments of the same nature, Mazar et al. (2008) concluded that self-concept deterring mechanisms, such as commitment devices, were an effective tool to stimulate truth-telling. While this conclusion was reached through a laboratory experiment, where lies concerned performance rather than morally questionable behaviour, the previously discussed literature, and the nature of the self-concept maintenance mechanism points towards its applicability in stimulating truth-telling in general as well.

2.2.3 Justifiability of behaviour

When it comes to the specific case of truth-telling, John et al. (2012) found that honesty also depends on the justifiability of the behaviour people are asked to report. Indeed, they found that people are naturally more honest when they believe their behaviour to be more justifiable. Since their survey experiment guaranteed the anonymity of their respondents, the main reason for these differences caused by justifiability may be self-concept-related. Indeed, it may be that people have a harder time admitting to behaviour that has a low degree of justifiability, since such an admission might naturally stimulate cognitive dissonance.

More generally, Zeelenberg and Pieters (2007) argued that the justifiability of a certain action greatly influences people's regret. They found that when there is a discrepancy between people's intentions and behaviour (i.e. cognitive dissonance), the behaviour is typically hard to justify, and this amplifies regret regardless of whether the action had a positive or a negative final outcome. Thus, when a certain behaviour is less justifiable, it takes a greater toll on the individual's self-concept (in this case, by amplifying regret).

Beyond the impact of justifiability on the way people react to a certain situation, the authors also recognised the decision process leading up to the action as the main determinant of justifiability (Zeelenberg & Pieters, 2007). Reb and Connolly (2010) developed this by stating that when people assess the consequences and value of an action more carefully – that is, go through a longer decision-process – they typically find the action more justifiable, and regret it less, once again regardless of the final outcome. This is based on the idea that justifiability relies on the nature of the action itself. When the individual considers whether to act or not to act more carefully, the resulting decision will reflect the individual's perception of the action better, such that it is the most justifiable of the options. At the same time, making a justifiable decision will reduce the probability that the individual blames himself (Reb & Connolly, 2010).

To this point, it is important to note that while the self-concept relies simultaneously on social and personal norms, justifiability can be based on either, depending on the situation and the individual. In other words, it is possible that someone finds a certain action justifiable while this same action is not justifiable in the eyes of society (Reb & Connolly, 2010). This further confirms how justifiability relies on the decision process, as a longer consideration of the action may provide the individual with more personal justifications for it.

All of these findings are in line with John et al. (2012), as they indicate that people will admit more frequently to actions with a high degree of justifiability, as these actions involve less cognitive dissonance. Conversely, people should be more reluctant to admit to more impulsive, less justifiable actions, although people may still try to increase the justifiability of their behaviour after they made their decision. As such, the individual's perceived justifiability for a certain action may still increase after the act as he finds ex-post reasons for behaving the way he did (Zeelenberg & Pieters, 2007).

2.3 Truth-telling Mechanisms

The commitment mechanism described above has been used in an attempt to stimulate truth-telling in previous laboratory and experiment settings, namely in performance reporting (Mazar et al., 2008), and in overclaiming questionnaires (Weaver & Prelec, 2013). Across these studies, it was compared to various other truth-telling mechanisms, producing mixed results. In performance reporting, the honour code virtually eliminated all lies (Mazar et al., 2008). In the overclaiming questionnaires, however, it was found to be less effective than, most notably, different versions of the Bayesian Truth Serum (Weaver & Prelec, 2013).

All of these experiments and studies focused on unambiguous behaviour, where subjects were only confronted with the decision of whether to lie or not. Furthermore, neither of these studies was able to create fully anonymous settings due to their respective setups. In the present study, the commitment mechanism will be tested in the context of reporting various types of unverifiable unethical behaviour. Given the rather unique nature of this type of data, the survey using an honour code will be compared to one using a Bayesian Truth Serum, as this was already proven to be an effective way to acquire unverifiable data (John et al., 2012).

The next section will give an overview of how the Bayesian Truth Serum (BTS) is able to stimulate truth-telling.

2.3.1 The Bayesian Truth Serum

The BTS is a survey format used to gather unverifiable data in a way that incentivises respondents to tell the truth. It revolves around the assumption that people are Bayesian updaters, meaning that they update their beliefs using all the information that is available to them. Furthermore, it is assumed that they act optimally according to everybody else's actions, thus acting towards a Bayesian Nash Equilibrium. This creates a situation in which the optimal action for the respondent is to tell the truth, as long as other respondents tell the truth as well (Prelec, 2004).

The BTS is based on a scoring method that relies on two questions. The first question requires respondents to report how they behave in a certain context, and the second asks for their estimation of how common this behaviour is among other people. The method yields high scores for "surprisingly common" outcomes, which occur when someone

overestimates how common their own behaviour is in the population. With this method, truth-telling is expected to systematically result in higher scores, thus incentivising respondents to tell the truth (Prelec, 2004).

The idea is that people tend to overestimate the frequency of their behaviour in the population (Prelec, 2004). This is derived from the "false consensus effect", which states that people overestimate the degree to which others are like them (Taylor, Peplau & Sears, 1994). Since they behave as Bayesian updaters, people use their knowledge of their own behaviour as a "sample of one", which they subsequently use to estimate other people's behaviour when they do not have access to other data (Dawes & Mulford, 1996). Prelec (2004) combines this with the "common prior" assumption to build the theoretical model that supports the BTS.

The common prior assumption states that all individuals have access and are aware of the same "common knowledge" regarding the population's behaviour. In other words, this means that two individuals that behave in the same way have exactly the same beliefs about how the rest of the population behaves (Prelec, 2004). Conversely, the beliefs of two people that behave differently only differ in the information that these people infer from their own behaviour, which comes down to the "sample of one" effect described previously (Dawes & Mulford, 1996).

Concretely, this assumption implies that when there are only two possible behaviours, resulting in two types of people, there will be a single belief regarding the population's behaviour for each type. Hence, all people that share a certain behaviour will have the same belief regarding this behaviour in the population.

Both aspects of the BTS score result from this distinction between the beliefs of people behaving in one way and those of people behaving in the other way. On the one hand, the type of the individual determines his or her information score (I_r), such that this information score is higher when the reported behaviour is more common than is predicted by the average respondent. On the other hand, the individual's prediction itself determines the prediction score (P_r) of the BTS. The resulting score for respondent r is computed as such:

$$T_r = I_r + P_r$$

The prediction score is computed such that the total score of the respondent (that is, T_r) is always highest when the respondent predicts the exact frequency of the relevant behaviour. According to the theory, since the respondent bases his prediction on the common prior and on his own behaviour only, and since he has no reason to believe that his prediction is wrong, this implies that his score should be higher if he is honest about his own behaviour than if he lies about it (Prelec, 2004).

The scoring method, as described previously, is then coupled to a form of monetary incentive. For instance, John et al. (2012) use a charity incentive to stimulate truth-telling by letting their respondents choose between various charities that will receive a donation depending on the respondent's score, where higher scores led to larger donations.

Weaver and Prelec (2013) recognise that the success of the BTS relies on two main factors: first, the respondents must believe that the scoring method does indeed reward truthful answers. Therefore, the survey should be fully transparent regarding the scoring method, and describe it sufficiently to convince the respondent of its validity. Second, the incentive must be greater than the cost of telling the truth. In John et al. (2012), the respondent must therefore care enough about the charity, and the donation should be large enough to make telling the truth the utility maximising response. In the present paper, the survey using the BTS relies on a similar incentive to John et al. (2012), while also using a similar question design.

2.3.2 Evidence of the effectiveness of the BTS (relative to other truth-telling mechanisms)

Psychological studies have shown that it was possible to stimulate truth-telling by making people believe that they were being exposed to a "monitoring device" that supposedly functioned as a lie detector (Jones & Sigall, 1971). As long as subjects believed in the ability of the machine to detect lies, this would deter them from doing so. Overtime, truth-telling mechanisms have become increasingly complex, using methods ranging from answer scoring (Prelec, 2004; Miller et al., 2005) to the exploitation of psychological characteristics (Mazar et al., 2008).

When it comes to the different scoring methods, studies have agreed that the subject's belief that truth-telling is the utility-maximising action is central to the effectiveness of

the relevant method. This of course applies to the BTS as well (Prelec, 2004; Barrage & Lee, 2009; Weaver & Prelec, 2013).

The BTS has been used in various settings since its introduction by Prelec (2004). For instance, Weaver and Prelec (2013) used it in the context of an overclaiming questionnaire, while Barrage and Lee (2009) applied it in a laboratory setting in which subjects were asked to state their preferred donations in different situations. One of the common denominators across both of these studies was that the data was collected in a controlled environment.

Weaver and Prelec (2013) provided their respondents with customised questionnaires, which they completed in a laboratory-like environment. Furthermore, the nature of the questionnaire allowed the authors to pinpoint the exact extent to which their subjects were lying on certain questions. This allowed them to provide clear conclusions with respect to the magnitude of people's dishonesty, and thus also on the absolute effectiveness of the BTS. Even though they focused primarily on the information score, rather than on the result of the complete BTS scores, the experiment indicated that the BTS, while it did not fully eliminate dishonesty or careless answering, was an effective way of stimulating truth-telling (Weaver & Prelec, 2013).

Barrage and Lee (2009) applied the BTS in a laboratory setting where the truth was not always observable to the experimenter, but relative increases in truth-telling were still measurable. While the BTS was found to stimulate truth-telling to a certain extent, it was not deemed to be the most effective truth-telling mechanism. Rather, its effect was found to be comparable to that of cheap talk. The authors attributed this to the fact that some of their subjects may not have understood the BTS mechanism, resulting in the inference that truth-telling would maximise their score to be seen as cheap talk.

From the previously mentioned studies, it is clear that the BTS can serve as an effective truth-telling mechanism, but also that it is subject to certain limitations that may cause other truth-telling mechanisms to yield superior results in different situations. In laboratory settings, when the questions do not impact the respondent's self-concept, the BTS was proven to work (Weaver & Prelec, 2013).

However, truth-telling is a major concern in studies that do not use laboratory experiments as well. John et al. (2012) applied the BTS to an anonymous online survey,

and found significant differences in truth-telling between the control survey and the BTS version. Furthermore, their study gave an indication with regards to the applicability of the BTS to potentially dissonance-arousing questions.

Similarly to the BTS, the Honour code as used by Mazar et al. (2008) was applied in various laboratory settings, yielding encouraging results. When compared to the BTS specifically, in the context of Weaver and Prelec's (2013) study, it appeared to reduce dishonesty to a much smaller extent than the monetary incentive. However, the authors recognised that their results might be different if, for instance, subjects were guaranteed anonymity.

While John et al. (2012) did not compare the BTS to any other truth-telling mechanism, their setting guaranteed the full anonymity of the respondents. However, this guarantee also implied that it was harder to verify exactly how well respondents understood the BTS scoring method, possibly affecting its effectiveness.

A major advantage of the Honour code resides in this uncertainty. Indeed, while it was shown to be less effective in controlled, laboratory settings, the respondent does not need to be convinced of how Honour code works, since it directly affects his perceived utility by creating cognitive dissonance (Mazar et al., 2008). The trade-off between experimental control and subject anonymity may thus greatly affect the relative effectiveness of the considered truth-telling mechanisms.

3. Hypotheses development

In the present study, the goal is to stimulate truth-telling by using a commitment reminder inspired by the experiments of Mazar et al. (2008) and Weaver and Prelec (2013). This commitment reminder will be compared to a BTS incentive, which is applied to the same questionnaire, and to a control questionnaire that does not involve any truth-telling mechanisms.

From the previously discussed literature, a number of hypotheses are formulated. First of all, the role of the Honour code in arousing cognitive dissonance should imply that people who are faced with such an Honour code will find it more difficult to lie while maintaining their self-concept. Because of this, it can be expected that people will be more

inclined to tell the truth when exposed to the Honour code survey. This leads to the first hypothesis:

H1a: The Honour code survey results in higher admission rates than the control survey.

Furthermore, the BTS incentive significantly improved truth-telling in previous survey studies that used similar dissonance-arousing questions (John et al., 2012; Beznea et al., 2016). Since the design of the present survey closely resembles that of these previous studies, expectations are that the BTS will lead to higher admission rates than the control survey as well. Hence, the same hypothesis is formulated for the BTS:

H1b: The BTS incentive results in higher admission rates than the control survey.

These two first hypotheses assess whether both truth-telling mechanisms actually work in the context of this study. Beyond this, the goal is also to determine which of these two mechanisms is the most effective in answering each question.

Previous findings are somewhat ambiguous in this regard, as each of these methods showed various degrees of effectiveness depending on the study and on the context in which they were used (Weaver & Prelec, 2013; Barrage & Lee, 2009; Mazar et al., 2008). Therefore, the second hypothesis is formulated in the null form:

H2: The admission rates between the BTS incentive and the Honour code survey do not differ.

The previously stated hypotheses constitute the primary focus of this study. However, in addition, the justifiability of every studied behaviour is assessed, and the truth-telling rates between each survey will also be compared on the basis of this justifiability index. While John et al. (2012) found a link between justifiability and truth-telling, their results were too context-specific to formulate an independent hypothesis on this aspect of the study. The resulting analysis will be discussed as an extension of the main hypotheses.

4. Survey Design

To test the previously mentioned hypotheses, a survey composed of six main questions regarding questionable behaviour was designed. This was done on the basis of previous research by John et al. (2012), as well as a study conducted by some colleagues and myself earlier this year (Beznea et al., 2016). Both of these studies used a control and a BTS

treatment that were in all respects identical. The present study follows a similar approach, in that the survey is designed such that all respondents answered the same questions, which were all presented in the same way, except for the addition of a question relating to the BTS incentive in one version of the survey.

On top of this, for the purpose of this research, a third version of the survey was created, where the BTS incentive was replaced by the online equivalent of an honour code, following Weaver and Prelec's (2013) design. Hence, three different versions of the same survey were randomly distributed amongst participants.

The following sections will describe all the common and specific features of each version of the survey. First, I discuss the core of the survey, after which I give a brief overview of the BTS and self-concept features.

4.1 Control survey

On the first page of the survey, respondents are informed that, in order to thank them for their participation, €100 will be donated to the following three charities: The World Wildlife Fund, Doctors Without Borders, The Dutch Cancer Society. This is done to make sure that the BTS incentive, which will make use of this donation, is not confounded by any form of "charity effect", which might otherwise bias responses in an unpredictable way. Along with this message, emphasis is put on the fact that the survey is fully anonymous, meaning that it is impossible to trace any of the answers provided in the survey back to the respondent (see Appendix A1.1).

Then, respondents are asked four general information questions (Appendix A1.2) which regard age, gender (male/female), nationality (Dutch/other) and current occupation (professional/ student/other). These questions are asked to get a general impression of the composition of the sample, and to assess overall representativeness of the results.

After these general questions, he or she is presented with the survey's six core questions. These questions regard questionable or unethical behaviour that people do not typically talk about. They were chosen specifically to apply to virtually any adult that would be susceptible to answer the survey, and such that they may provoke various degrees of dishonesty amongst respondents. The questions are:

- *Have you ever **intentionally** walked out of a supermarket/store without paying for one (or more) items?*
- *Have you ever **intentionally** ignored recycling norms by throwing something in the wrong bin?*
- *Have you ever cheated on someone you were/are in a romantic relationship with?*
- *Have you ever lied in a job, school or internship application to improve your chances?*
- *Have you ever lied to someone you just met to appear more interesting/attractive?*
- *Have you ever lied about the death or health condition of a relative as an excuse for your own bad behaviour?*

For each of these questions, the respondent is presented with four options to answer: Never, Once or twice, Occasionally, and Frequently (Appendix A1.3). This scale was previously used in John et al. (2012) and the study conducted by my colleagues and I (2016). In both instances, it proved to be adequate in discerning the effects of the BTS incentive on the respondents' answers, both in terms of admission rates and frequency of the relevant behaviour.

It is important to note that the questions themselves were designed carefully to make sure that respondents who intentionally misreport their behaviour will only do so in one direction. That is, respondents may understate the frequency of a certain behaviour, but have no reason to overstate it. This is crucial in order to be able to discuss how admission rates are influenced by each version of the survey.

Furthermore, it may be the case that respondents answer these questions differently depending on which one they see first, or depending on how many they have already answered. Because of this, the order in which these questions appear for each respondent is randomised, to avoid potential order effects, and the respondent cannot change his answer to a question once he has moved on to the next one.

Each of the main questions is followed by a question where respondents are asked to predict what percentage of the other respondents they think have behaved according to what the previous question asked at least once (Appendix A1.3). This question is asked for every one of the studied behaviours. This set of questions will be used subsequently to calculate the BTS score of every respondent that answers the survey that uses the BTS incentive. To ensure that all three versions of the survey are as similar as possible,

however, these questions are presented to every respondent, regardless of which version he or she is answering.

Answers to the prediction questions make use of sliders that range from 1 to 99 (percent) to provide the respondent's answer. This is done to make sure that all respondents use the same scale (i.e. percentages rather than proportions) and that they all answer within the range of valid responses. This choice follows from the confusion experienced by some subjects during the study conducted by my colleagues and I (2016), where responses were recorded both in percentages and proportions, making the data transformations unnecessarily complicated at later stages of the analysis. In addition, the values of 0 and 100 percent are left out, since these values are incompatible with the BTS scoring method, and have no practical meaning, in the sense that it cannot be assumed that the entire population displays perfectly homogeneous behaviour.

The final set of questions (Appendix A1.4) regards the perceived justifiability of each of the studied behaviour. Once again, the respondent is faced with six questions, regarding whether he or she believes that the relevant behaviour is justifiable. Following the set-up from John et al. (2012), respondents are presented with three choices (Yes/Possibly/No, to the question of whether something is justifiable).

Eliciting the justifiability of these behaviours will enable deeper comparative analysis around the main questions and how truth-telling around each of the studied behaviours is influenced by the different versions of the survey. Although this is not backed up by any form of previous research, the main questions were designed such that they are expected to result in various degrees of justifiability.

4.2 BTS incentive

The version-specific questions for each survey are placed after the questions on general information (Appendix A1.2). This is done to maximise the effect of the truth-telling incentives on the respondents' answers on the main questions.

The BTS incentive is formulated in similar fashion to the research my colleagues and I conducted previously. In the spirit of John et al.'s (2012) set-up, a donation to charity is used as a monetary incentive. For obvious reasons, direct personal monetary incentives cannot be used while guaranteeing the respondent's anonymity. Hence, the subjects responding to the BTS version of the survey are exposed to an additional question after

they complete the general information section. This additional question asks them to choose one of three charities that they would like to donate money to. The charities are chosen such that they represent very diverse causes, thus maximising the probability that each respondents cares about at least one of them.

The choice of charity is coupled with a message that explains the workings of the BTS scoring mechanism. Respondents are told that maximising their BTS score will ensure that more money goes to their preferred charity, as the money will be allocated based on how their individual score compares to the average score of all respondents, thus resulting in a linear relationship between the individual's BTS score and their contribution to their preferred charity. As stressed by Weaver and Prelec (2013), emphasis is put on the fact that telling the truth is expected to yield systematically higher BTS scores, such that respondents are aware that truth-telling should maximise the share of the money that goes to their preferred charity.

The phrasing of the message itself is based on the survey that was set up by my colleagues and I (2016), as the study found significant differences in truth-telling between control and BTS surveys, indicating that the respondents understood and were responsive to the content and framing of the message.

4.3 Self-concept statement

Finally, the third version of the survey uses an honour code that is placed after the general information questions, just like the BTS incentive was. As such, both versions of the survey are identical in all respects, except for the content of the addition question.

For the honour code, the message was inspired by the "solemn oath" used by Weaver and Prelec (2013), with the difference that, to guarantee anonymity, the name of the respondent is not included in the oath itself. The message is presented as follows:

> *"Before starting the questionnaire,* ***please certify that you will answer all of the subsequent questions truthfully*** *by agreeing to the following statement:*
>
> *I promise to answer all the questions of this survey truthfully."*

The respondents are made to "sign" this honour code by clicking on the statement, making sure that they not only read it, but also actively engaged in agreeing to it. This

should make them more consciously aware of their desire to be honest, and thus increase cognitive dissonance in case they decide to lie in the remainder of the survey.

5. Data

The survey was conducted online from the 11th to the 29th of May 2017. During this time, over 1600 emails were sent to first, second and third-year Economics and Business students from the Erasmus University of Rotterdam. The same email was sent to a couple hundred staff members as well. The phrasing of the email can be found in Appendix A2.

In addition, the survey was posted on various social media groups specific to various studies at the EUR, including groups from the Erasmus MC. The message that was used to direct respondents to the survey was in all respects identical to the emails.

The following sections will describe the general characteristics of the collected data, as well as an overview of the descriptive statistics of the main questions. On top of this, I provide some basic insights into the answers of the main questions.

5.1 General characteristics

Over the course of the experiment, a little less than 800 respondents started one of the versions of the survey. Of these responses, approximately 50 were discarded, as these respondents did not answer any of the main questions. The final dataset thus makes use of a total of 736 observations. Of these, 711 were complete, meaning that the respondent answered every single question of the survey, 4 failed to answer the questions on justifiability, but answered all the main questions, and 21 answered at least one of the main questions.

Respondents were randomly assigned to one of the three treatments (control, BTS or self-concept). The control group is composed of 242 observations, of which 237 are complete, and 5 are partial. The BTS group is composed of 247 observations, including 9 partial responses, and the self-concept group is composed of 247 observations, including 11 partial responses.

Information on *age*, *gender*, *occupation* and *nationality* were derived for all considered observations. In addition, the answers to the main questions resulted into six categorical variables about degrees of admission to the relevant questionable behaviour, which are generally referred to as *shoplifting*, *recycling* (ignoring recycling norms), *cheating*,

application (lying on a job, school or internship application), *attractive* (lying to appear more interesting/attractive), and *death* (lying about the health condition or death of a relative). These variables are also transformed into binary variables that simply distinguish whether people admitted to having behaved a certain way at least once or not. Discussion about admission rates refers to the binary version of these variables, while discussion about degrees of admission refers to the initial categorical variables. The prediction questions resulted in the *estimated percentage* of each of these behaviours amongst the other respondents. Finally, the *justifiability* of each behaviour was computed on the basis of all complete observations.

The average age across treatments is 21.7, and ranges from 21.6 for the control group to 21.9 for the self-concept group. The youngest respondent reported to be 16, while the oldest reported an age of 68. Furthermore, 93% of the full sample are students. This is 95% in the control group, and 92% for both the BTS and the self-concept group.

The sample is made out of 48% males, and 52% females. This distribution is almost identical across all treatments. Finally, 57% of the sample is of Dutch nationality, which is slightly overrepresented in the BTS sample, where 61% of respondents is Dutch.

Table 1: General information

Variable	**Obs**	**Mean**	**Std. Dev.**	**Min**	**Max**
Age	736	21.71	4.51	16	68
Gender (1=Male, 0=Female)	736	0.48	0.50	0	1
Nationality (1=Other, 0=Dutch)	736	0.43	0.50	0	1

5.2 Descriptive statistics

When it comes to the behaviours that the respondents are asked about in the main questions, important differences are observed both in terms of justifiability and in terms of general admission rates. Justifiability is computed such that it can take on any value from 0 to 2, where 0 implies that none of the respondents found the behaviour justifiable at all, and 2 means that every respondent found the behaviour completely justifiable. Cheating and lying about the death or health condition of a relative appear to have a very low degree of justifiability, using the same computation as John et al. (2012), with scores of 0.32 and 0.34, respectively, across 711 observations. On the other hand, lying on a job

or school application, or lying to appear more attractive, show high levels of justifiability (0.86 and 1.00, respectively). Table 2 shows the justifiability scores for every behaviour.

Table 2: Justifiability scores

Questionable behaviour	***Obs***	***Justifiability***
Cheating	711	0.32
Lying about the death of a relative	711	0.34
Shoplifting	711	0.45
Ignoring recycling norms	711	0.54
Lying on application	711	0.86
Lying to appear more attractive	711	1.00

Across the entire sample, admission rates generally follow a similar pattern. Cheating and lying about the health condition or death of a relative present the lowest admission rates, with 27% and 29%, respectively. Although it appears to be less justifiable, people admit to ignoring recycling norms more than to lying on a school or job application (68% against 47%). Besides, admission rates are highest for lying to appear more attractive, as 76% of respondents admit to this, irrespective of which treatment they are in. An overview of admission rates across and within treatments is presented in Figure 1.

When taking a closer look at the admission rates in each sample, it can be seen that both the BTS and the self-concept treatment yield higher admission rates than the control treatment for every one of the observed behaviours (Figure 1). Furthermore, it appears that respondents in the latter two treatments admit to more questionable behaviour, either by admitting to higher frequencies of one behaviour, or by admitting to more different behaviours. This is confirmed by computing a total admission score (TAS) for each respondent. This TAS is obtained by adding up the numerical values associated to each answer given to the six main questions (where “Never” is 0, “Once or twice” is 1, “Occasionally” is 2, and “Frequently” is 3). The resulting average TAS for respondents in the control group was 3.41, while it was 4.88 and 4.73 for the BTS and the self-concept treatment respectively.

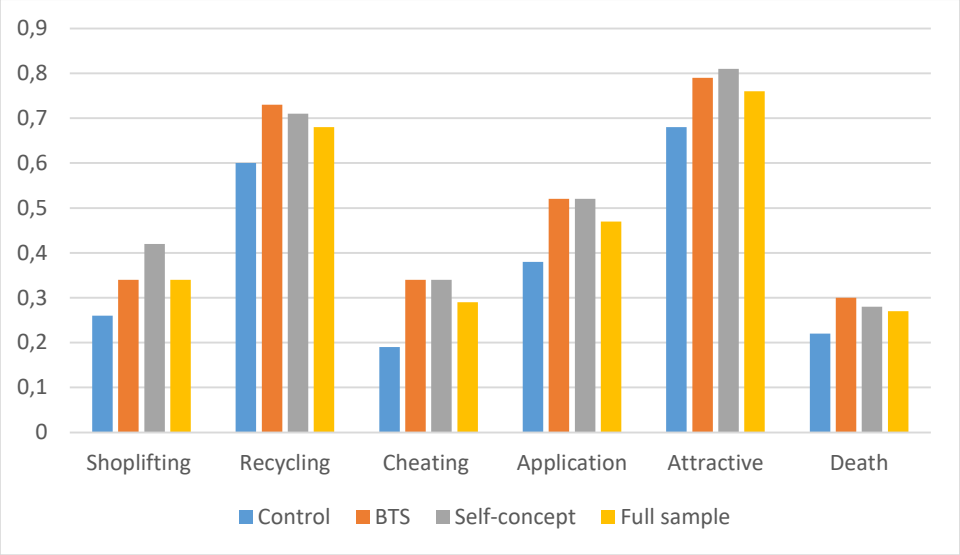

Figure 1: Admission rates across treatments

More specific distributions of admission rates and frequencies for each of the treatments can be found in Appendix A3. Generally, all possible answers to each question were fairly well-represented, as at least 5 observations can be counted for each frequency, with the exception of the number of "frequent" admissions to shoplifting in the control group (2) and to lying about the health condition or death of a relative in every treatment (0, 4 and 1 in control, BTS and self-concept, respectively).

Other low counts of observations are found in "frequent" admissions to cheating (5) in the control group, and to lying on job or school applications in the control and the self-concept groups (6 and 8, respectively). The latter may be due to the fact that the large majority of respondents were young students that have not had the chance to complete many such applications yet.

Finally, the estimates resulting from the prediction questions are summarized in Figure 2. Although theory predicts that there should be no significant differences between the observed means of estimates across treatments, the means recorded in the BTS and in the self-concept treatment are systematically slightly higher than those in the control group. Similarly, means in the self-concept treatment are generally higher than in the BTS treatment. These differences may be due to the fact that respondents may expect less lying in the BTS and self-concept versions of the questionnaire.

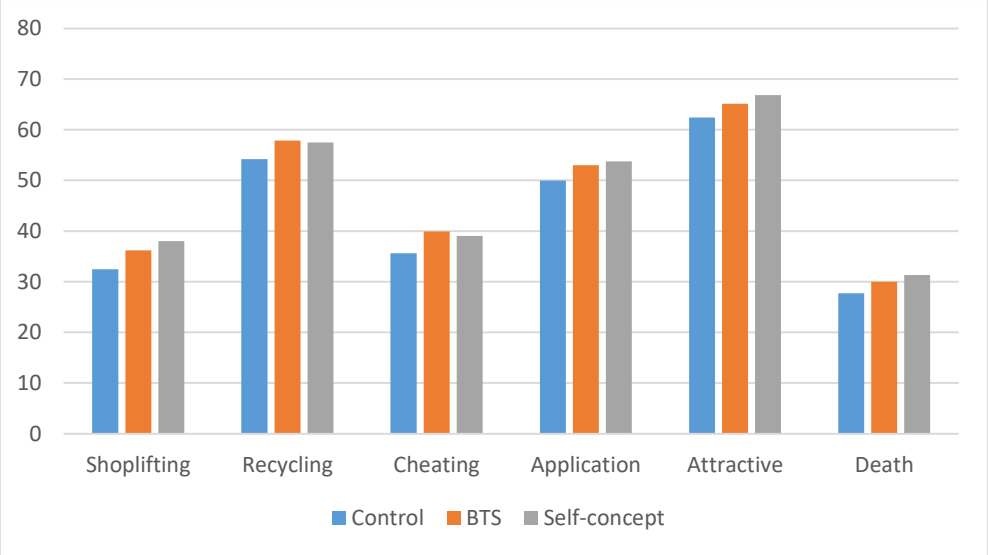

Figure 2: Descriptive statistics for prediction questions

6. Methodology

In order to analyse the experimental data and discuss the hypotheses in more detail, I perform a series of comparative tests. The following sections detail which tests are used to test the validity of each hypothesis, then cover a couple of additional tests that can be performed on the dataset to gain additional insights in the truth-eliciting mechanisms.

6.1 Hypothesis 1

Hypothesis 1 is divided into two parts that compare both treatments that involve truth-elicitation mechanisms to the control group. As such, both parts of the hypothesis are answered by using the same tests. Initially, each of the six main questions will be covered individually, and the admission rates in the control group will be compared to those in the BTS and in the self-concept group, respectively. For this comparison, two-sample chi-squared tests are used. In addition, the Bonferroni correction will be applied to the critical significance levels that are used, in order to account for the testing of multiple hypotheses (one for each of the main questions) simultaneously.

In accordance with the hypothesis, the expectation is that respondents admit to each questionable behaviour significantly more often in the two treatments involving truth-eliciting mechanisms relative to the control group. The chosen test will allow not only to determine whether the difference between treatments is significant, but also to discuss the direction of this difference.

Subsequently, a series of Mann Whitney U tests are run to determine whether there are significant differences in the degree of admission to all researched behaviours. This second series of comparative tests is performed based on the finding of John et al. (2012), who found that asking respondents about different degrees of admission also affected the respondents' honesty with regards to the degree to which they admit to a certain behaviour. Together, these tests provide sufficient insights to reach reliable conclusions regarding both parts of the first hypothesis.

However, in order to check for the validity of the truth serum question, a t-test is applied to every prediction question to check whether the average *estimated percentage* of other respondents behaving questionably is significantly different between people who admit to the relevant behaviour and people who do not. As was highlighted in the related literature on the BTS incentive, this is a necessary condition for the incentive to be successfully implemented. These tests will be discussed prior to answering the hypotheses, in order to determine to what extent the common prior assumption holds.

6.2 Hypothesis 2

Hypothesis 2 aims to compare both truth-eliciting mechanisms and their relative effectiveness. For this, a similar methodology to that of hypothesis 1 is used. The main difference comes from the fact that no clear expectations were formulated as to which treatment was most effective. In case significant differences are found, the two sample chi-squared and the Mann Whitney U test will provide further insight regarding the direction of these differences.

Furthermore, the tests are still performed on both the binary and on the categorical version of the variables, to discuss possible differences in degrees of admission between both treatments as well.

6.3 Additional tests

Besides the main comparative tests, a logistic regression is performed on the full sample of complete observations for each question, using the binary variables of the main questions as the dependent variable, and the treatments as an independent categorical variable. This would also provide additional insights on the effect of *age, gender, occupation* and *nationality* on the admission to each of the questionable behaviours.

The marginal effect of the treatments in these regressions corresponds to the observed admission rates in each treatment.

The final regression that is performed requires the data to be reshaped into panel data. This is done such that the answer to every question becomes a single observation, meaning that six observations are recorded for every respondent. In order to look at how the admission rate across all behaviours was dependent on the different variables, the binary admission variables are pooled together as a single variable, which is used as the dependent one in the regression. This enables the inclusion of question dummies as independent variables, to see how the answer to these questions can explain part of the variation in overall the admission rate.

7. Results

This section starts by covering the results of the control tests on the prediction questions, in order to confirm that the common prior assumption was indeed satisfied in the BTS treatment. Next, I present the analysis of the results for both hypotheses. This is followed by a more general discussion of the findings of the experiment.

7.1 BTS assumption

In order to insure that the common prior assumption (Prelec, 2004) indeed holds across the responses collected in the BTS group, a two-sample t-test is performed for every question, comparing the mean estimated percentage of other respondents behaving according to the relevant behaviour given by people that admit to the behaviour, and people that don't. Finding consistently higher means amongst the people that admit to the behaviour would confirm that the BTS incentive can be properly applied to the questions in this survey experiment.

The statistical analysis confirms that, in the BTS group, respondents that admit to a certain behaviour also estimate a frequency of this behaviour amongst other respondents that is significantly higher than the frequency that is, on average, estimated by the respondents that do not admit to this behaviour. This holds for every question, and all the differences in means are significant at a 1% level (p-value = 0.0000 for all behaviours, see appendix A4), confirming the validity of the BTS incentive.

Although the difference in distribution is significant for every question, the common prior assumption theoretically implies that there should be only two beliefs in the population: one for people that admit to the behaviour, and one for the people that do not. To further verify to what extent this is true, the estimates given by the respondents in the BTS groups are investigated. Since estimates for both types of people in each question show similar standard deviations, the individual values are observed to see what share of each distribution is more likely to be part of the other.

To do this, each value is compared to the mean of its own distribution, as well as to the mean of the second distribution. Across all questions, it is observed that about 24% of all estimates given by the respondents in either category are closer to the mean of the other category. These deviations are as low as 18% for the question regarding ignoring recycling norms, and as high as 30% for cheating. Furthermore, these common prior "errors" seem to occur as much when respondents that do not admit to a certain behaviour overestimate the frequency of the behaviour as when those who admit underestimate it.

The fact that the errors do not appear to be systematically linked to one of the two categories of respondents seems to point towards the imperfection of the common prior assumption rather than to any design issues in the survey. It indicates that although the distributions are sufficiently different to insure that the BTS assumptions are generally satisfied, truth-telling is only the utility maximising outcome for a majority of the respondents, rather than for all of them. It may thus be that a respondent whose answer violates the common prior assumption may be better off lying, depending on the actual outcome of the prediction score.

7.2 Hypothesis 1

As was mentioned in the discussion of descriptive statistics, all six behaviours generally show higher admission rates in the BTS and self-concept treatments compared to the control. To give an overview of the relative increases in admission rates, the odds ratios between each of these treatments and the control group are displayed for each of the six behaviours in table 3, in similar fashion to John et al. (2012). It can be noted that the relative increases in admission rates differ quite substantially between behaviours, going as low as 16% in the BTS group, for lying to appear more attractive, and as high as 77% for cheating on a romantic partner, in the BTS group as well.

When comparing the BTS to the control group, significant differences in admission rates are found in four of the six questions when applying the Bonferroni correction for the simultaneous testing of multiple hypotheses. More specifically, the questions regarding ignoring recycling norms, lying on an application, and lying to appear more attractive were significant at a 5% level, while the difference in admission about cheating was found to be significant at a 1% level. Shoplifting and lying about the death of a relative only presented significant differences before applying the Bonferroni correction, at a 10- and 5% level, respectively.

A similar pattern is observed when looking at the comparison between the self-concept and the control group. Here, five of the six questions show significant differences, with ignoring recycling norms being significant at a 10% level, while lying on an application and lying to be more attractive display significance levels of 5%. Just like in the comparison between the BTS and control groups, cheating on a romantic partner shows significant differences at a 1% level. In addition, respondent seem to admit significantly more to shoplifting under this treatment as well, as the difference is significant at a 1% level. Lying about the death of a relative is the only behaviour that does not show significant differences, even without taking the Bonferroni corrections into account. An overview of the significance levels is given in table 3.

Table 3: Odds ratios for admission rates between treatments

Variable	BTS/Control ratio	Self-concept/Control ratio
Shoplifting (1=yes, 0=no)	1.28*	1.59***(***)
Ignoring recycling norms (1=yes, 0=no)	1.22***(**)	1.18**(*)
Cheating (1=yes, 0=no)	1.77***(***)	1.75***(***)
Lying on application (1=yes, 0=no)	1.36***(**)	1.35***(**)
Lying to appear more attractive (1=yes, 0=no)	1.16***(**)	1.18***(**)
Lying about the death of a relative (1=yes, 0=no)	1.38**	1.26

Note: significance levels of two-sample chi-squared tests are 0.1, 0.05** and 0.01***. In parentheses, significance levels are shown when the Bonferroni correction is applied to the critical levels, such that 0.1/6 = 0.017(*), 0.05/6 = 0.0085(**) and 0.01/6 = 0.0017(***). Odds ratios are computed by taking the admission rates in either treatment and dividing them by the admission rate in the control group.*

Overall, the findings on admission rates for the six main questions seem to unambiguously support hypotheses 1a and 1b, confirming the expectations that resulted from the conclusions of previous literature. The fact that only positive differences were

found is in line with John et al. (2012), who found effects in the same direction for nine of their ten questions on questionable research practices when comparing the BTS to a regular questionnaire. Furthermore, these findings hint at the fact that Mazar et al.'s (2008) conclusions regarding how stimulating people's self-concept to make them more honest may be applicable outside of laboratory settings.

It is interesting to note, however, that the significance of the differences in admission rates do not seem to be entirely driven by the perceived justifiability of the considered behaviour. John et al. (2012) expected greater differences in admission rates as the behaviour became less justifiable, however, this does not appear to be the case here, and the least significant results across both treatment comparisons are found for lying about the death of a relative, which, as seen in the section 5.2, was the least justifiable behaviour, alongside cheating on a romantic partner. It is all the more interesting to note that, in turn, cheating shows the most significant differences across comparisons. Although it is difficult to assess with certainty where these results come from, they may be explained by the different nature of these behaviours. This may lead to differences in the way they are perceived, yielding low degrees of justifiability for different reasons. This provides further evidence that the effects of the BTS – and possibly other truth-telling mechanisms – may be dependent on the context in which it is applied, and on the questions that are asked.

Similarly, shoplifting only shows strongly significant differences when comparing the self-concept to the control group, while displaying low to no significance levels in the BTS comparison. This is perhaps most surprising with regards to the findings in the self-concept group. Indeed, the magnitude and significance levels found between the BTS and control group are comparable to those found in the study conducted by my colleagues and I earlier this year (Beznea et al., 2016), while the self-concept treatment yielded much higher admission rates than were previously found. Further discussion on the possible implications of this finding are discussed in section 7.3, in the context of the second hypothesis.

Looking at the Mann-Whitney U tests that were performed on the categorical variables for each question, the results once again confirm John et al.'s (2012) observations. It appears that respondents in the BTS treatment not only admit significantly more to the

relevant behaviour; they also admit to higher frequencies of said behaviour. Still applying the Bonferroni correction, differences between the BTS and the control group are significant at a 5% level for ignoring recycling norms, and at a 1% level for cheating, lying on an application, and lying to appear more attractive. Shoplifting and lying about the death of a relative are still not significant under the critical levels imposed by the Bonferroni correction, however, both of their p-values decreased relative to those from the tests on the binary variables (from 0.0801 to 0.0172 and 0.0368 to 0.0308 without Bonferroni correction, respectively), indicating that more important difference in distribution were observed when considering different degrees of admission. This decrease in p-values can explained by the fact that the categorical variables give a more precise overview of the changes in admission rates from one treatment to the other.

When comparing the self-concept treatment to the control, significant differences in degrees of admission are observed for lying on an application at a 5% level, and for shoplifting, cheating and lying to appear more attractive at a 1% level. Lying about the death of a relative still doesn't display any significance. It is worth noting that ignoring recycling norms no longer presents significant differences in degrees of admission under the Bonferroni correction, while they were significant when looking at admission rates.

Furthermore, all the rank statistics revealed positive associations between the two treatments and the control. John et al. (2012) argued that this may be due to the fact that the framing of the answers as various frequencies may induce respondents to lie not only about the admission itself, but also about how frequently they behave as is asked in the question. The evidence provided here seems to confirm this assessment.

7.3 Hypothesis 2

While the discussion of the first hypothesis emphasised clear differences between the control and the two truth-telling stimulating treatments, these treatments also need to be compared to each other, in order to determine whether one is more effective than the other. The odds ratios of the admission rates between the BTS and the self-concept treatment are once again computed, and are displayed in table 4. The first observation is that this time, the numbers revolve around 1.00, with four of the six main questions resulting in odds ratios between 0.98 and 1.03. In addition, there does not seem to be a clear trend of either treatment resulting in higher admission rates across all questions.

Analysis of the binary variable for admission rates confirms that there are no significant differences between the two treatments for any of the main questions, under the Bonferroni correction. When looking at the critical values before subjecting them to the correction, only shoplifting shows significant differences between the two treatments, and only at a 10% level. It appears that respondents admit significantly more to shoplifting under the self-concept treatment. As was already touched upon in the previous section, this is all the more surprising since a study on similar behaviour had found rates closer to the ones found in the BTS group.

It seems unlikely that the self-concept treatment indeed triggered higher truth-telling rates for this particular question, while it appears to result in virtually identical rates for all other questions. A possible explanation, however, could reside in the nature of the behaviour that respondents are asked about. Shoplifting could be considered as the only behaviour that is unambiguously criminal, and it may be the case that the treatments are better at stimulating truth-telling on different kinds of behaviour.

In line with this, the only notable difference in admission rates, although not significant, can be found in lying about the death of a relative, where the BTS results in 10% more admissions than the self-concept survey. In this case, the behaviour is not formally illegal, but is heavily looked down upon. The nature of both incentives – one monetary and one stimulating cognitive dissonance – could cause the respondents to react differently to them depending on the question they are asked. This was already hinted at in previous literature (John et al., 2012; Weaver & Prelec, 2013).

Overall, however, the results of the comparisons on the binary admission variables strongly point towards the absence of differences between the BTS and the self-concept treatments.

Table 4: Odds ratio of admission rates between BTS and self-concept treatments

Variable	BTS/Self-concept ratio
Shoplifting (1=yes, 0=no)	0.80*
Ignoring recycling norms (1=yes, 0=no)	1.03
Cheating (1=yes, 0=no)	1.01
Lying on application (1=yes, 0=no)	1.01
Lying to appear more attractive (1=yes, 0=no)	0.98
Lying about the death of a relative (1=yes, 0=no)	1.10

Note: significance levels of Mann Whitney U tests are 0.1, 0.05** and 0.01***. None of the comparisons show significant differences under the Bonferroni correction.*

When considering the variables as categorical, the conclusions drawn from the binary variables are confirmed and strengthened, as shoplifting no longer shows significant differences between treatments without the Bonferroni correction. Hence, both treatments show similar distributions when it comes to the degree of admission for every behaviour. This indicates that the conclusions that were drawn by John et al. (2012) regarding how respondents react to the BTS incentive when given the choice between various degrees of admissions can probably be applied to the self-concept treatment as well. That is, respondents treat both incentives similarly, in the sense that they are effective in stimulating truth-telling not only in the context of binary admission rates, but also in making respondents more truthful regarding the frequency of their behaviour.

Although no clear expectations were formulated regarding the second hypothesis, the null stating that there are no significant differences in admission rates between the BTS and the self-concept treatment is confirmed for both binary and categorical variables. This is sharply contrasting with the conclusions reached by Weaver and Prelec (2013), who found their solemn oath treatment, which is equivalent to the self-concept one used in this study, to be largely ineffective in stimulating truth-telling, while the BTS significantly did.

7.4 Logistic regressions

Besides the comparative non-parametric tests that were used to answer the hypotheses, logistic regressions were performed for each of the studied behaviours, to study the effect of the various controls that were gathered on the admission rates for the relevant behaviour. The treatment variable was included as the main independent variable as well, in order to check whether the conclusions draw from the non-parametric tests still held

after including the available controls. The coefficients for every variable in each of the six regressions are displayed in table 5.

Looking at the significance of the treatment variables confirms the conclusions drawn in the previous two sections, as the coefficients of both treatments show the same level of significance as in the non-parametric tests. More interesting are the values and significance of the remaining variables. The first thing to note is that occupation does not seem to be significant in any of the regressions. This is not surprising, as over 93% of the full sample was composed of students, limiting the variability that could be observed as caused by differences in occupation.

Age does not appear to influence the admission rates of five of the six behaviours, as all coefficients are insignificant and close to 0, except in the regression that concerns admission rates to cheating on a romantic partner, where it is highly significant (at a 1% level under the Bonferroni correction), and positively related to admission rates. A possible reason for this could be that old respondents have been in more romantic relationships, and thus had more opportunities to cheat than younger respondents. This is especially likely as the average age was 21, making it quite conceivable that some of the younger respondents have never been in a romantic relationship before.

Gender did not have a significant impact on four of the six studied behaviours, however, it is worth noting that five of the six coefficients are positive, implying that generally, male respondents admit to the studied behaviour more than females. This relationship is significant for two of the behaviours, which are shoplifting (at a 5% level), and lying to appear more attractive (at a 5% level, and at a 10% level using the Bonferroni correction).

Finally, respondents were asked whether they were Dutch or not, and results reveal that non-Dutch respondents admit significantly more to shoplifting, cheating and lying on an application (at 5-, 1- and 1% significance levels under the Bonferroni correction, respectively). On the other hand, they admit less to lying to appear more attractive, although this result is less significant (5% without Bonferroni correction) than the previous three. These findings may be due to cultural differences, although this is beyond the scope of this research.

Table 5: Logistic regressions for each binary admission variable

Variables	(1)	(2)	(3)	(4)	(5)	(6)
Treatment						
BTS	0.37*	0.62***(**)	0.87***(***)	0.61***(***)	0.55***(*)	0.45**
Self-concept	0.71***(***)	0.50**(*)	0.79***(***)	0.56***(**)	0.65***(**)	0.32
Age	0.01	0.03	0.09***(***)	-0.01	-0.02	0.01
Male	0.33**	0.14	0.16	-0.11	0.44**(*)	0.08
Occupation						
Student	-0.38	0.55	-0.17	-0.33	0.39	-0.23
Professional	-0.48	-0.19	-0.45	-0.46	1.11	-0.60
Non-Dutch	0.45***(**)	-0.10	0.80***(***)	0.63***(***)	-0.41**	0.13
C	-1.32	-0.80	-3.73***(***)	-0.16	0.69	-1.30
R^2	0.0311	0.0166	0.0777	0.0290	0.0307	0.0077

Note: The coefficients for the logistic regressions are displayed using the following binary variables as dependent variables: (1) Shoplifting; (2) Ignoring recycling norms; (3) Cheating; (4) Lying on an application; (5) Lying to appear more attractive; (6) Lying about the death of a relative. Significance levels are 0.1, 0.05** and 0.01***. In parentheses, significance levels are shown when the Bonferroni correction is applied to the critical levels, such that 0.1/6 = 0.017(*), 0.05/6 = 0.0085(**) and 0.01/6 = 0.0017(***).*

Finally, it is worth noting that despite the previously mentioned significance levels, all the regression models display fairly low R^2 levels, indicating the low explanatory power of these models (<10% for all six regressions).

As discussed in the section 6.3, the final regression concerns the overall admission rate when each answer to one of the main questions is considered as an independent observation. The coefficients and significance levels resulting from this regression are displayed in table 7.

The findings obtained from this panel regression are in line with those obtained previously. Unsurprisingly, both treatments are strongly significant (at a 1% level), which indicates that the effects found across most of the individual questions are also visible over the experiment in general. This confirms that both treatments are in general effective in stimulating truth-telling in the context that was defined by the survey used in this study.

On top of this, the question dummy displayed significant coefficients for every question. As these coefficients are indexed on one of the questions, the coefficients can be interpreted as the relative proportion of admissions between questions. As such, the

behaviours that were admitted to most (ignoring recycling norms and lying to appear more attractive) show the highest coefficients, while the coefficients of cheating and of lying about the death of a relative are negative, as these behaviours were less common than shoplifting, which serves as the benchmark.

Table 6: Panel logistic regressions for binary admission variable (with clustered std. err.)

Variables	(1)
Treatment	
BTS	0.56***
Self-concept	0.58***
Age	0.02*
Male	0.15*
Occupation	
Student	-0.07
Professional	-0.26
Non-Dutch	0.27***
Question	
Recycling	1.44***
Cheating	-0.24**
Application	0.57***
Attractive	1.86***
Death	-0.37***
C	-1.58***
$\mathbf{R^2}$	0.1253

Note: The coefficients for the logistic regressions are displayed using the overall binary admission variable as the dependent variable, first without including the estimates as independent variables (1), then incorporating these in the model (2). The question variable is treated as a categorical variable, hence no coefficient is shown for shoplifting, as this serves as the benchmark for the other questions. Significance levels are 0.1, 0.05** and 0.01***.*

Controls regarding age and gender are marginally significant. The only control that stands out is the one regarding nationality, as being non-Dutch seems to be associated with significantly more admission, at a 1% level.

In terms of explanatory power, this model performs a lot better than the regressions found in table 5, which is most likely due to the inclusion of question-specific variables in a model that groups all questions into one panel set. This does suggest that admission

rates for a specific question are not completely dissociated from the admission rates for the other questions.

The same panel logistic regression was performed including an interaction term between the treatment and the question variables, however, these interactions did not result in significant coefficients, and did not alter the conclusions that were reached using the previous model. Because of this, this regression was left out of the analysis. However, it indicates that there may not be question specific treatment effects, which is an interesting observation.

8. Discussion

The following sections will discuss various possible extensions and limitations around the approach and conclusions to this study. First, I deal with the main limitations around the design of the incentives, after which I give an overview of the general limitations surrounding the data collection. These sections, 8.1 to 8.2, specifically highlight possible mechanisms leading to overestimation of the admission rates. Section 8.3 addresses a conceptual issue with the self-concept incentive. Section 8.4 discusses what would happen when some of the most extreme observations are excluded from the sample, and why this may be worth considering. Furthermore, some points regarding the value and magnitude of the findings are addressed in sections 8.4 and 8.5. Finally, some suggestions are made for further analysis.

8.1 Common prior assumption

In setting up a BTS incentive, the common prior assumption plays a key role in determining whether it can be successfully implemented. Although evidence was found of significant differences in the distribution of the respondents' beliefs depending on what they answered, analysis of the individual answers also revealed that over a fifth of the responses presented clear violations of this common prior assumption.

Although it is unclear to what extent this jeopardised the effectiveness of the BTS, it clearly hints at the possibility that truth-telling was not the utility maximising choice for some of the respondents, thus introducing a risk that the BTS incentive unintentionally deceived respondents into believing that they were better off telling the truth. While it is unlikely that this played a role in driving the results, it is a possibility that cannot be excluded, given that results supported the conceptual effectiveness of the BTS.

The main concern in this case arises because it may imply that respondents no longer have reasons to lie in only one direction (that is, denying a certain behaviour instead of admitting it). Indeed, the possibility exists that respondents, were they aware of this issue with the common prior assumption, may have found lying in either direction to be utility maximising. It is easy to imagine that while admitting to a behaviour is self-concept deterring when a respondent is actually guilty of such behaviour, it would be virtually costless for the respondent to admit to a certain behaviour if he or she did not actually behave this way, as long as the gain in utility from lying exceeds the mere cost of lying.

This bides the question of whether there is a possibility that the BTS actually overestimated the actual admission rates. Indeed, if respondents underestimate behaviour overall, then respondents that should not admit to the behaviour, but that are aware of this underestimation now have an incentive to lie by admitting to the behaviour. A comparison of estimates and admission rates in the BTS treatment indicates that that this is true for two of the six behaviours, namely *ignoring recycling norms* and *lying to appear more attractive*.

8.2 Sample and selection biases

A second point of concern regards the sample of respondents that was used. The population of respondents that filled in the various versions of the survey largely came from the student population of the Erasmus University of Rotterdam. This is the same population that was used for the previous study that was conducted by my colleagues and I (Beznea et al., 2016). This study had already used the BTS in an almost identical design to the present one. As the results of this previous study had been discussed in the campus magazine at the beginning of the calendar year of 2017, many students were informed about the goal of the study, and the means used to reach the findings.

Although it is impossible to verify exactly to what extent the present sample overlaps with the one that was used in the previous study, as anonymity was guaranteed, it may be problematic that some respondents were aware not only of the workings of the BTS incentive, but also of the general design of the experiment. This is another factor that may have contributed to the overestimation of certain admission rates, in the BTS group, but also, to a certain extent, in the self-concept group. Unfortunately, it is impossible to control for this phenomenon, given that collecting data outside of the Erasmus University would not have been feasible in terms of reaching similar numbers of respondents.

Regarding respondent selection, another obvious concern regards possible self-selection biases. Indeed, respondents were reached through various online platforms, as was explained in section 5, however, the people reached were under no obligation to engage with the survey they were provided with. As such, it is worth remarking that the large majority of students that were reached were Dutch (about 2/3 of the emails sent, and 4/5 of the students that were engaged through other online platforms). Despite this, only 57% of respondents were Dutch, indicating that international students may have been overrepresented in the final study.

Finally, in terms of selection, it is important to note that, given the nature of the six main questions, it may have been the case that some respondents had simply not been in contact with the relevant behaviour, meaning that they would not even have been susceptible to admit to it at all. This is not necessarily a problem in terms of comparing the two truth-telling treatments to the control group, however, it may have been the case that it made some questions irrelevant for some respondents, thus somewhat biasing the obtained admission rates.

In addition, the average age of respondents also affected the frequency of behaviour they admitted to, as the same number of "offences" may have been relatively less important for older people than for students. This may challenge the generalisability of the conclusions.

8.3 Conceptual issue of the Honour code

Despite the apparent effectiveness of the self-concept incentive, a potential problem of using the Honour code in the context of increasing truth-telling on questionable behaviours is that it compensates for being honest regarding bad behaviour – a process that should increase cognitive dissonance and thus possibly hurt self-concept – by creating another form of cognitive dissonance. As such, it attempts to compensate for a negative impact on the subject's self-concept by making people aware of the benefits of honesty, which would be self-concept enhancing. Predicting the effectiveness of the Honour code on truth-telling in such situations thus depends on how past and present actions influence self-concept.

This concern can be addressed using past literature that was mentioned earlier in this paper. It was found that the perceived justifiability of a certain behaviour makes it less

self-concept deterring, even when said behaviour is questionable (John et al, 2012; Reb & Connolly, 2010). Furthermore, Zeelenberg and Pieters (2007) showed that the justifiability of an action can still increase after the action was performed.

This last finding implies that an action is likely to be more justifiable when looking at it retrospectively, as compared to when the subject is faced with this same action in the present. Hence, the fact that people have more opportunities to find justifications for their past actions would make thinking of these actions less self-concept deterring than if they were to perform these actions again in the present.

This appears to be supported, albeit in a different context, by Ramsøy's (2015) basic model of consumer choice. This model uses a neuroscientific approach to split choice and decision-making into various stages of experience. In particular, he recognises a predicted and an experienced value before and at the time of the decision, but also a remembered value after the decision. Crucially, he argues that the remembered value is subject to change depending on subsequent experiences and thoughts.

The way we remember a certain action, and the value we associate to this action is not fixed overtime. When relating this to the conclusions of Greenwald (1980) and Sanitioso et al. (1990), who showed that people tend to associate themselves with more desirable qualities or outcomes, this provides substantial evidence that the recollection of past behaviour should result in less deterrence of a subject's self-concept than a decision they are presently confronted with.

Hence, it would appear that past and present actions are perceived differently, and thus, the cognitive dissonance generated by an Honour code should dominate the cognitive dissonance of being confronted with a past instance of questionable behaviour. This seems to be supported by the findings of my study.

8.4 Extreme TAS scores

Given the concerns that were previously raised about the familiarity of the sample with the design of the study, and about the possibility of overestimation in either of the two treatment groups, a closer look was taken at the collected observation. Using the previously mentioned *total admission score* (TAS, see section 5.2), it was noted that the 26 observations presenting the highest TAS all belonged to the BTS and the self-concept group. While this is not necessarily surprising, these scores (between 13 and 16 out of a

possible 18) were remarkably high. In comparison, the highest score for an individual in the control group was 12, and this was only found for a single individual.

Out of curiosity, and to verify the robustness of the findings of the main study, the same analysis and tests were ran excluding all observations that showed TAS scores above 12. As a result, 17 observations were taken out of the BTS, and 9 out of the self-concept. The fact that nearly two thirds of these observations belonged to the BTS group comforts the doubts around whether this treatment in particular was subject to overestimation of admission rates or degrees of admission.

Although the exclusion of these observations obviously reduced significance of the observed differences between treatments (all excluded individuals admitted to at least five of the six behaviours), all tests showed similar results to the ones performed on the full sample. Four of the six behaviours still present significantly higher admission rates in the BTS treatment compared to the control group, and this also holds for five of the six behaviours in the self-concept treatment.

Perhaps the only notable difference is that results now seem to point towards the self-concept treatment being slightly more effective than the BTS treatment overall, as it yields equal or higher admission rates for five of the six behaviours.

This, of course, only holds if all 26 observations are indeed the result of overestimations. The assumption is still that the majority of these observations is valid. Hence, conclusions should be drawn on the basis of the information provided by the full sample. However, the additional tests ran on the alternate sample strengthen these conclusions by confirming that both treatments are still effective even when removing the more extreme observations.

8.5 Relative effect of BTS and self-concept treatments

One of the goals of the study was to compare the effectiveness of the BTS to that of the self-concept treatment. Findings generally indicated that these treatments had very similar effects on the respondent's admission rates across five of the six questions. However, given the concerns for overestimation in the BTS group, and the fact that no real expectations were formulated around the comparison of these two treatments, the conclusions drawn regarding the second hypothesis should be taken with a grain of salt.

Both treatments bring along different types of limitations. While those around the common prior assumption were discussed in section 8.1, the self-concept incentive also presents its own set of limitations. For instance, it may be argued that the nature of the questions asked in the survey already impacts the self-concept regardless of the incentive. As such, it may be that this self-concept incentive only offsets the adverse effect of the behaviour in question on the respondent's perception of himself.

This may be all the more problematic given that three of the questions are directly asking about lying, which is a form of dishonesty, and which therefore comes down to exactly what the incentive is supposed to prevent. Although this ambiguity was discussed in section 8.3, no clear assessment was made about the degree to which both lies (the past one that the respondent is asked about and the present one would he fail to admit to a behaviour he is guilty of) interact. This may cause confounding effects in some questions, while leaving others unaffected, thus biasing the results of the experiment.

Furthermore, the self-concept incentive was based on commitment devices that had previously only been used in field experiments or laboratory settings, where the respondent had to physically sign some form of "oath" (Weaver & Prelec, 2013). This study was the first of its kind to compare this type of commitment device to a BTS treatment in an online survey. This may have led to other unforeseen effects, as it may have reduced the tangibility of the commitment. Although the self-concept treatment was successful in yielding higher admission rates than the control group, this limitation further draws into question the reliability of the comparison between the BTS and the self-concept treatment.

8.6 Link to previous literature and incentive effects

As was mentioned in the discussion of the results, findings are in line with most of the directly related previous literature that looks at truth-telling or honesty (John et al., 2012; Mazar et al., 2008). The present findings confirm the idea, which had been put forward in psychology and marketing research, that monetary incentives are not the only feasible stimulants of truth-telling among survey respondents specifically.

However, the only other study that had compared the BTS to a solemn oath, which the self-concept incentive was based on, reached different conclusions regarding the

effectiveness of both of these treatments. Weaver and Prelec (2013) found the BTS to be far superior to their solemn oath when it came to truth-telling about verifiable metrics.

Although the crucial differences between the present study and the one of Weaver and Prelec (2013) were discussed alongside the rest of the related literature, it is somewhat surprising that the present study finds virtually no differences between the two treatments. One possible explanation for this may be that both treatment were perceived in a similar way by the respondents.

If this were the case, it could be that most respondents did not take the BTS as a monetary incentive, but rather as an explicit reminder to tell the truth. This would make the BTS incentive very similar to the self-concept one in terms of what it accomplishes, which is to stimulate cognitive dissonance and thus impact the respondent's ability to maintain his self-concept while lying.

As such, it could be the case that the observed effects in this research were the result of a single factor rather than of two distinct ones. This would explain the very small differences in admission rates between the two treatments, as it implies that they would then both stimulate the same response.

A reason for this may be the size of the incentive, which was substantially smaller than the previous one offered by my colleagues and I (2016). It may have been the case that this amount of money was too small for respondents to consider the BTS incentive seriously. The similarities in results for both treatments then arise from the fact that the BTS may stimulate self-concept while also adding an incentive on top of this. If the incentive is ineffective, the remaining effect could be attributed to the self-concept stimulant.

If such a pattern were to be observed in reality, then Weaver and Prelec's (2013) estimation that the BTS is more effective than their solemn oath might also hold in the setting provided in the present study. Further analysis using the same questions with more salient monetary incentive are required to draw more solid conclusions in this regard.

8.7 Link between admission and predictions

Another matter that creates opportunity for further research is the link between the respondent's estimate in the prediction question and whether or not he admits to a certain behaviour. For endogeneity concerns, this relationship could not simply be tested using regressions similar to those presented in section 7.4, however, results provided around the common prior assumption suggest that such a relationship exists.

Regressions performed on the full sample including the estimates as well as all controls as independent variables returned significant positive coefficients for these estimates on the admission rate to every question. However, given the previously mentioned concerns of endogeneity, these regressions are not taken into consideration in the discussion of the main results of this study. Instead, they serve as an indication that the previously mentioned links may be worth investigating.

Further research in this area could focus on determining whether the estimate can be used to determine the likelihood that an individual will admit to a certain behaviour, or on how much the admission to a certain behaviour influences the individual's estimate, effectively trying to determine the importance of the "sample of one" effect in influencing the individual's prediction.

In the context of the present study, it may also be interesting to see how the effect of the estimates influences the effect of the treatments, and whether the inclusion of these estimates in a robust model would in any way alter the conclusions that were reached in the previous sections.

8.8 Justifiability

The final point of discussion regards the metric of justifiability, which was not developed extensively in the present study. The concept of justifiability and its relationship with the observed admission rates to certain behaviour offers a major avenue for further research. As shown in multiple studies around the BTS and truth-telling incentives (John et al., 2012; Beznea et al., 2016), there may exist a clear link between the odds ratio of treatment over control and the justifiability of a certain behaviour.

This relationship was already made evident in the literature on self-concept (Zeelenberg & Pieters, 2007; Reb & Connolly, 2010), however, little has been done to determine how justifiability affects both absolute and relative levels of admission. Additional research

could focus on using this metric to determine when different truth-telling mechanisms would be most effective when it comes to eliciting more accurate admission rates on questionable behaviour.

It may also be interesting to see whether a justifiability index could be created across different types of behaviour. One can think of combining the present study with that of John et al. (2012), and to further expand the list of behaviours by adding different types of questionable actions either in general, or in specific contexts.

9. Conclusion

The aim of the study was to compare two truth-telling mechanisms to a baseline control and to each other. All in all, the results of the analysis were quite conclusive. Significant differences in admission rates were found between the control and the two truth-telling treatments. It was determined that both mechanisms significantly increased truth-telling rates amongst respondents. No significant differences were found between the two mechanisms themselves.

Both the BTS and the self-concept incentive appear to be effective in increasing truth-telling amongst respondents. Although the nature of the questions did not allow to check to what degree it eliminated lying, these result present a valuable contribution to the already existing research on eliciting the unverifiable truth.

Economic applications for these findings could range from gathering more trustworthy information on employee ethics and behaviour to customer moral hazard in the context of insurance. At a research level, it may be a valuable stepping stone to gain more insight on what motivates people to tell the truth. As such, further research could expand on this in various fields. In economics and psychology specifically, a similar setup could be created in a laboratory setting, or with a different sample of non-student respondents to see whether the present findings can be replicated in more natural settings. Furthermore, one could think of applying a similar methodology to different questions and contexts.

Bibliography

Aronson, E. (1969). The theory of cognitive dissonance: A current perspective. *Advances in experimental social psychology*, *4*, 1-34.

Aronson, E. (1980). Persuasion via self-justification: Large commitments for small rewards. *Retrospection on social psychology*, 3-21.

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

Beznea, A., Harding, D., Huang, Y., van Hulsen, M., Schneider, L. S. (2016). *Shoplifting and the Bayesian Truth Serum* (working paper). Erasmus School of Economics.

Bok, D. C. (1990). *Universities and the Future of America*. Duke University Press.

Campbell, E. Q. (1964). The internalization of moral norms. *Sociometry*, 391-412.

Cohen, A. R., Greenbaum, C. W., & Mansson, H. H. (1963). Commitment to social deprivation and verbal conditioning. *The Journal of Abnormal and Social Psychology*, *67*(5), 410.

Cunha, M. P. E., & Cabral-Cardoso, C. (2006). Shades of gray: A liminal interpretation of organizational legality-illegality. *International Public Management Journal*, *9*(3), 209-225.

Dawes, R. M., & Mulford, M. (1996). The false consensus effect and overconfidence: Flaws in judgment or flaws in how we study judgment?. *Organizational Behavior and Human Decision Processes*, *65*(3), 201-211.

De Quervain, D. J., Fischbacher, U., Treyer, V., & Schellhammer, M. (2004). The neural basis of altruistic punishment. *Science*, *305*(5688), 1254.

Dickerson, C. A., Thibodeau, R., Aronson, E., & Miller, D. (1992). Using cognitive dissonance to encourage water conservation. *Journal of Applied Social Psychology*, *22*(11), 841-854.

Freedman, J. L. (1965). Long-term behavioral effects of cognitive dissonance. *Journal of Experimental Social Psychology*, *1*(2), 145-155.

Gergen, K. J. (1965). Effects of interaction goals and personality feedback on the presentation of self. *Journal of Personality and Social Psychology*, *1*(5), 413.

Gibson, R., Tanner, C., & Wagner, A. F. (2013). Preferences for truthfulness: Heterogeneity among and within individuals. *The American Economic Review*, *103*(1), 532-548.

Gneezy, U. (2005). Deception: The role of consequences. *The American Economic Review*, *95*(1), 384-394.

Greenwald, A. G. (1980). The totalitarian ego: Fabrication and revision of personal history. *American psychologist*, *35*(7), 603.

Griffin, D. W., & Ross, L. (1991). Subjective construal, social inference, and human misunderstanding. *Advances in experimental social psychology*, *24*, 319-359.

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.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: behavioral experiments in 15 small-scale societies. *The American Economic Review*, *91*(2), 73-78.

Hsee, C. K., Yu, F., Zhang, J., & Zhang, Y. (2003). Medium maximization. *Journal of Consumer Research*, *30*(1), 1-14.

Ickes, W. J., Wicklund, R. A., & Ferris, C. B. (1973). Objective self-awareness and self-esteem. *Journal of Experimental Social Psychology*, *9*(3), 202-219.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 0956797611430953.

Jones, E. E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin*, *76*(5), 349.

Lundquist, T., Ellingsen, T., Gribbe, E., & Johannesson, M. (2009). The aversion to lying. *Journal of Economic Behavior & Organization*, *70*(1), 81-92.

Maas, V. S., & Van Rinsum, M. (2013). How control system design influences performance misreporting. *Journal of Accounting Research*, *51*(5), 1159-1186.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of marketing research*, 45(6), 633-644.

McCabe, D. L., & Trevino, L. K. (1993). Academic dishonesty: Honor codes and other contextual influences. *The Journal of Higher Education*, *64*(5), 522-538.

McCabe, D. L., Trevino, L. K., & Butterfield, K. D. (2002). Honor codes and other contextual influences on academic integrity: A replication and extension to modified honor code settings. *Research in higher Education*, *43*(3), 357-378.

Miller, N., Resnick, P., & Zeckhauser, R. (2005). Eliciting informative feedback: The peer-prediction method. *Management Science*, *51*(9), 1359-1373.

Pennebaker, J. W., Hughes, C. F., & O'heeron, R. C. (1987). The psychophysiology of confession: Linking inhibitory and psychosomatic processes. *Journal of Personality and Social Psychology*, *52*(4), 781.

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

Ramsøy, T. Z. (2015). *Introduction to neuromarketing & consumer neuroscience*. Neurons Inc..

Reb, J., & Connolly, T. (2010). The effects of action, normality, and decision carefulness on anticipated regret: Evidence for a broad mediating role of decision justifiability. *Cognition and Emotion*, *24*(8), 1405-1420.

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A neural basis for social cooperation. *Neuron*, *35*(2), 395-405.

Sanitioso, R., Kunda, Z., & Fong, G. T. (1990). Motivated recruitment of autobiographical memories. *Journal of Personality and Social psychology*, *59*(2), 229.

Sartre, J. P. (1956). Being and Nothingness: An Essay on Phenomenological Ontology. *Trans. Hazel E. Barnes. New York: Philosophical Library*.

Schweitzer, M. E., & Hsee, C. K. (2002). Stretching the truth: Elastic justification and motivated communication of uncertain information. *Journal of Risk and Uncertainty*, *25*(2), 185-201.

Taylor, S.E., Peplau, L.A., and Sears, D.O. (1994). *Social Psychology* (8th ed). Englewood Cliffs, NJ: Prentice–Hall.

Walster, E. (1965). The effect of self-esteem on romantic liking. *Journal of Experimental Social Psychology*, *1*(2), 184-197.

Walster, E. (1970). The effect of self-esteem on liking for dates of various social desirabilities. *Journal of Experimental Social Psychology*, *6*(2), 248-253.

Weaver, R., and Prelec, D., (2013). Creating Truth-Telling Incentives with the Bayesian Truth Serum. *Journal of Marketing Research, Vol. L* (June 2013), 289–302.

Zeelenberg, M., & Pieters, R. (2007). A theory of regret regulation 1.0. *Journal of Consumer psychology*, *17*(1), 3-18.

APPENDIX A1: Survey outline

A1.1: First page

Dear respondent,

For my master thesis, I am researching responses to certain types of questionable behaviour. For this purpose, I put together a survey around five questions regarding situations anyone could be confronted with at any point in their lives. Note that the survey is completely **anonymous** and will be used for research purposes only.

To thank you and all other respondents for your participation, I will be donating a total of €100 spread across these three charities:

1. The World Wildlife Fund (for information about the charity, click here)
2. Doctors Without Borders (for information about the charity, click here)
3. The Dutch Cancer Society (KWF Kankerbestrijding) (for information about the charity, click here)

Thank you for your participation!

Luc Schneider

A1.2: Preliminary questions

Q2 What is your age? ___

Q3 What is your gender?

- ❍ Male (1)
- ❍ Female (0)

Q4 What is your current occupation?

- ❍ Student (1)
- ❍ Professional (2)
- ❍ Other (0)

Q5 What is your nationality?

- ❍ Dutch (0)
- ❍ Other (1)

[BTS incentive – For BTS version only]

Q6 As a result of this survey, I will be donating €100 to the following three charities. **You can influence how much your preferred charity will receive by answering all questions honestly!**

I use an objective scoring method developed by MIT professor Prelec to evaluate your answers. Using this method, honest answers are given a higher score. You can find a more detailed explanation of the scoring method here: Prelec, 2004. After the study, I will donate the money according to how your score compares to the scores of the other respondents. Providing honest answers will thus maximise the donation that will go to the charity of YOUR choice.

Keep in mind that the survey is completely anonymous and will be used for research purposes only.

Choose the charity you want us to donate to:

- ❍ The World Wildlife Fund (1)
- ❍ Doctors Without Borders (2)
- ❍ The Dutch Cancer Society (KWF Kankerbestrijding) (3)

[Honour code – For self-concept version only]

Q7 Before starting the questionnaire, please certify that you will answer all of the subsequent questions truthfully by agreeing to the following statement:

- ❍ I promise to answer all the questions of this survey truthfully. (1)

A1.3: Main questions and prediction questions

Q9 Have you ever **intentionally** walked out of a supermarket/store without paying for one (or more) items?

❍ Never (0)

❍ Once or twice (1)

❍ Occasionally (2)

❍ Frequently (3)

Q15 What percentage of other people do you think ever **intentionally** walked out of a supermarket/store without paying for one (or more) items?

_____ % (0)

Q10 Have you ever **intentionally** ignored recycling norms by throwing something in the wrong bin?

❍ Never (0)

❍ Once or twice (1)

❍ Occasionally (2)

❍ Frequently (3)

Q16 What percentage of other people do you think ever **intentionally** ignored recycling norms by throwing something in the wrong bin?

_____ % (0)

Q11 Have you ever cheated on someone you were/are in a romantic relationship with?

❍ Never (0)

❍ Once or twice (1)

❍ Occasionally (2)

❍ Frequently (3)

Q17 What percentage of other people do you think ever cheated on someone they were/are in a romantic relationship with?

_____ % (0)

Q12 Have you ever lied in a job or school application to improve your chances?

❍ Never (0)

❍ Once or twice (1)

❍ Occasionally (2)

❍ Frequently (3)

Q18 What percentage of other people do you think ever lied in a job or school application to improve their chances?

_____ % (0)

Q13 Have you ever lied to someone you just met to appear more interesting/attractive?

❍ Never (0)

❍ Once or twice (1)

❍ Occasionally (2)

❍ Frequently (3)

Q19 What percentage of other people do you think ever lied to someone they just met to appear more interesting/attractive?

_____ % (0)

Q14 Have you ever lied about the death or health condition of a relative as an excuse for your own bad behaviour?

❍ Never (0)

❍ Once or twice (1)

❍ Occasionally (2)

❍ Frequently (3)

Q20 What percentage of other people do you think ever lied about the death or health condition of a relative as an excuse for your own bad behaviour?

_____ % (0)

A1.4: Justifiability questions

Q19 Do you think that **intentionally** walking out of a supermarket/store without paying for one (or more) items can be justifiable?

❍ No (0)

❍ Possibly (1)

❍ Yes (2)

Q20 Do you think that **intentionally** ignoring recycling norms by throwing something in the wrong bin can be justifiable?

❍ No (0)

❍ Possibly (1)

❍ Yes (2)

Q21 Do you think that cheating on someone you were/are in a romantic relationship with can be justifiable?

❍ No (0)

❍ Possibly (1)

❍ Yes (2)

Q22 Do you think that lying in a job or school application to improve your chances can be justifiable?

❍ No (0)

❍ Possibly (1)

❍ Yes (2)

Q23 Do you think that lying to someone you just met to appear more interesting/ attractive can be justifiable?

❍ No (0)

❍ Possibly (1)

❍ Yes (2)

Q24 Do you think that lying about the death or health condition of a relative as an excuse for your own bad behaviour?

❍ No (0)

❍ Possibly (1)

❍ Yes (2)

APPENDIX A2: Invitation email

Dear [insert field of study] students,

Have you ever cheated on someone? Or have you ever shoplifted? And how many other people do you think have done these things before? I don't know about you, but I want to know!
You can help me out by filling in this survey on different kinds of questionable behaviour! It only takes 5 minutes, and I'll be donating €100 to charity as a thank you for your participation!

https://erasmusuniversity.eu.qualtrics.com/jfe/form/SV_8dZwKs7e38YWVox

In advance, thank you!

Luc Schneider,
MSc in Behavioural Economics
Erasmus School of Economics

APPENDIX A3: Distributions of admission rates within treatments

Variable	Control		BTS		Self-concept	
Shoplifting	Count	Prop.	Count	Prop.	Count	Prop.
Never	178	0.74	161	0.66	139	0.58
Once or twice	53	0.22	45	0.19	61	0.25
Occasionally	9	0.04	15	0.06	19	0.08
Frequently	2	0.01	22	0.09	21	0.09
Yes	64	0.26	82	0.34	101	0.42
No	178	0.74	161	0.66	139	0.58
Recycling						
Never	96	0.40	66	0.27	71	0.29
Once or twice	53	0.22	64	0.26	66	0.27
Occasionally	65	0.27	68	0.28	70	0.29
Frequently	25	0.10	46	0.19	34	0.14
Yes	143	0.60	178	0.73	170	0.71
No	96	0.40	66	0.27	71	0.29
Cheating						
Never	193	0.81	159	0.66	162	0.66
Once or twice	34	0.14	59	0.24	53	0.22
Occasionally	7	0.03	13	0.05	18	0.07
Frequently	5	0.02	10	0.04	11	0.05
Yes	46	0.19	82	0.34	82	0.34
No	193	0.80	159	0.66	162	0.66
Application						
Never	148	0.62	117	0.48	117	0.48
Once or twice	61	0.25	75	0.31	82	0.34
Occasionally	25	0.10	41	0.17	35	0.14
Frequently	6	0.03	11	0.05	8	0.03
Yes	92	0.38	127	0.52	125	0.52
No	148	0.62	117	0.48	117	0.48
Attractive						
Never	75	0.32	50	0.21	47	0.19
Once or twice	98	0.41	86	0.36	94	0.39
Occasionally	55	0.23	76	0.31	78	0.32
Frequently	10	0.04	30	0.12	23	0.10
Yes	163	0.68	192	0.79	195	0.81
No	75	0.32	50	0.21	47	0.19
Death						
Never	186	0.78	171	0.70	176	0.72
Once or twice	43	0.18	58	0.24	53	0.22
Occasionally	9	0.04	12	0.05	13	0.05
Frequently	0	0.00	4	0.02	1	0.00
Yes	52	0.22	74	0,30	67	0.28
No	186	0.78	171	0,70	176	0.72

APPENDIX A4: Common prior assumption

Table A4 1: T-test for significance of the difference in estimates between people that admitted to shoplifting and people that did not in the BTS group

Two-sample t-test with equal variances

Group	***Observations***	***Mean***
No admission	161	26.37 (1.395)
Admission	82	55.48 (2.209)
Difference		-29.11 (2.613)
t		-11.14
Degrees of freedom		148
Pr(|T| > |t|)		0.0000*

***p-value<0.01**

Notes: Difference = mean(no admission)-mean(admission), standard errors in parentheses.

Table A4 2: T-test for significance of the difference in estimates between people that admitted to ignoring recycling norms and people that did not in the BTS group

Two-sample t-test with equal variances

Group	***Observations***	***Mean***
No admission	66	27.42 (3.094)
Admission	178	69.12 (1.711)
Difference		-41.70 (3.536)
t		-11.79
Degrees of freedom		108
Pr(|T| > |t|)		0.0000*

***p-value<0.01**

Notes: Difference = mean(no admission)-mean(admission), standard errors in parentheses.

Table A4 3: T-test for significance of the difference in estimates between people that admitted to cheating and people that did not in the BTS group

Two-sample t-test with equal variances

Group	***Observations***	***Mean***
No admission	159	34.04 (1.393)
Admission	82	51.23 (2.200)
Difference		-17.19 (2.603)
t		-6.60
Degrees of freedom		148
Pr(|T| > |t|)		0.0000*

***p-value<0.01**

Notes: Difference = mean(no admission)-mean(admission), standard errors in parentheses.

Table A4 4: T-test for significance of the difference in estimates between people that admitted to lying on an application and people that did not in the BTS group

Two-sample t-test with equal variances

Group	***Observations***	***Mean***
No admission	117	38.44 (1.953)
Admission	127	66.37 (1.877)
Difference		-27.93 (2.709)
t		-10.31
Degrees of freedom		242
Pr(|T| > |t|)		0.0000*

***p-value<0.01**

Notes: Difference = mean(no admission)-mean(admission), standard errors in parentheses.

Table A4 5: T-test for significance of the difference in estimates between people that admitted to lying to appear more attractive and people that did not in the BTS group

Two-sample t-test with equal variances

Group	***Observations***	***Mean***
No admission	50	45.32 (2.950)
Admission	192	70.33 (1.484)
Difference		-25.01 (3.302)
t		-7.57
Degrees of freedom		77
Pr(|T| > |t|)		0.0000*

***p-value<0.01**

Notes: Difference = mean(no admission)-mean(admission), standard errors in parentheses.

Table A4 6: T-test for significance of the difference in estimates between people that admitted to lying about the death of a relative and people that did not in the BTS group

Two-sample t-test with equal variances

Group	***Observations***	***Mean***
No admission	171	23.80 (1.243)
Admission	74	44.41 (2.197)
Difference		-20.61 (2.525)
t		-11.79
Degrees of freedom		123
Pr(|T| > |t|)		0.0000*

***p-value<0.01**

Notes: Difference = mean(no admission)-mean(admission), standard errors in parentheses.