ERASMUS UNIVERSITEIT ROTTERDAM

ERASMUS SCHOOL OF ECONOMICS

How can the Bayesian Truth Serum help people to live a healthier lifestyle?

Student name:	K. (Karlijn) de Wilde BSc
Student ID number:	413155kw
Supervisor:	Dr. (Sophie) S.C. van der Zee
First reader:	M.A.J. (Merel) van Hulsen MSc
Master:	Economics and Business
Specialisation:	Behavioural Economics
Date:	12-16-2018

How can the Bayesian Truth Serum help people to live a healthier lifestyle?

Abstract

People tend to under-report food intake and over-report physical activity. The misreporting of healthy behaviour is a problem because this leads to an inaccurate assessment of the current state of healthy behaviour in studies. This paper examines whether the Bayesian Truth Serum (BTS) can be used to elicit more honest and thoughtful self-reported responses to healthy behaviour questions. In order to estimate the effectiveness of the BTS in eliciting honest self-reports, an experimental design questionnaire was developed. Two hundred and twenty-four Dutch females completed the questionnaire, which included questions about general eating behaviour, fruit consumption, vegetable consumption, general physical activity, moderate physical activity, and intensive physical activity. A 2-by-2 between subject design was used for this study. Subjects were asked whether they were actively trying to lose weight or not, and in each group subjects were randomly assigned to the BTS group or the control group. It was found that the BTS did not elicit more honest and thoughtful self-reported responses. In contrast, results suggest that the BTS did elicit less honest responses to questions about physical activity. Furthermore, respondents who are watching their weight are more likely to report healthy behaviour. However, it was not found that the effectiveness of the BTS differed among weight watchers or non-weight watchers. Overall, the BTS did not provide more accurate self-reports in this particular study of healthy behaviour.

Keywords: Bayesian Truth Serum, eating behaviour, physical activity, self-reports, over-reporting, under-reporting

Table of Contents

1 Introduction

In the Netherlands, in 2016, almost half of the population was classified as overweight (Leefstijlmonitor CBS, 2018). Moreover, the number of people who are overweight is rising, affecting the need for help in losing weight. Bunt, Mérelle, Steenhuis, and Kroeze (2017) found that 24.9 percent of the Dutch population would like to get help in losing weight, and 6.4 percent already got help in losing weight. These recent trends that more people are overweight and would like to get help in losing weight have heightened the need for methods to effectively help people to lose weight. Two proven actions an individual can take to increase weight loss are reducing food intake and increasing physical activity (Klesges, Klesges, Haddock, & Eck, 1992). However, taking the desired actions seems to be difficult to put into practice.

Although losing weight can be difficult it is important since being overweight is correlated with health risks (Leefstijlmonitor CBS, 2018). Therefore, it is crucial to increase the likelihood that people reduce food intake and increase physical activity. Previous research has demonstrated that people are more capable of changing their behaviour when they are more aware of what they eat and how much they exercise (Bandura, 1998). Self-monitoring is a way to become aware of eating and exercising behaviour. It is also proven to be an effective method to lose weight (Burke, Wang, and Sevick, 2011). However, there are also some limitations to self-monitoring methods. One well-studied limitation is the incorrectness of self-reports. Notably, several external factors might influence people to under-report food intake (Hill & Davies, 2001; Schoeller, 1990; Trijsburg et al., 2016) and to over-report physical activity (Rzewnicki, Auweele, and De Boudeaudhuij, 2003; Van de Mortel, 2008). For self-reporting to affect weight loss, it is essential that the self-reporting is done truthfully, consistently and timely (Bandura, 1998). When people misreport, behaviour self-monitoring methods might be less effective. The misreporting is also a problem since research investigating healthy behaviour can be misleading and wrong conclusions can be made. For example, a long time ago researchers believed that people with obesity ($BMI \geq 30$) did eat less than people with a healthy BMI. The researchers concluded that their obesity was due to a metabolic defect (Macdiarmid & Blundell, 1998). Later, it was found that this was not true. In contrast, obese people were more likely to under-report food intake and therefore the wrong conclusion was made. This example shows that it is crucial to have reliable measures of healthy behaviour.

Consequently, it is scientifically relevant to examine ways to reduce the under-reporting of non-healthy behaviour and make self-reports more truthful. Prelec (2004) invented a way to

induce truthtelling in questionnaires about behaviour. The method Prelec used is called the Bayesian Truth Serum (BTS) and it is used when the true behaviour cannot be known objectively or when it is hard to know the true behaviour. The BTS elicits more truthful answers by rewarding people to tell the truth. Respondents are rewarded with the BTS score, which is measured by their estimation of what others would do in a particular situation, in comparison to their own behaviour. In this paper, the BTS will be applied to the healthy behavioural setting in an attempt to elicit more truthful self-reports. This leads to the following research question:

Can the Bayesian Truth Serum elicit more honest and thoughtful self-reported responses to healthy behaviour questions?

The term healthy behaviour will be used in this paper to refer to desirable behaviour concerning someone's health. More specifically, healthy behaviour, in this context will refer to actions that lead to weight loss or maintenance of a healthy weight. The type of healthy behaviour used in this paper regards healthy eating behaviour, including fruit- and vegetable consumption, and healthy physical activity, including moderate- and intensive physical activity (Health Council of the Netherlands, 2017a; The Netherlands Nutrition Centre, n.d.). A healthy weight is defined by having a Body Mass Index (BMI) between 18.5 and 25 (The Netherlands Nutrition Centre, n.d.). Someone is overweight when she has a BMI of 25 and above. BMI is another factor that influences the misreporting of healthy behaviour. Previous research demonstrated that people with a higher BMI are more likely to under-report true non-healthy behaviour (Heitmann & Lissner, 1995). In addition, people that are trying to lose weight are also more likely to under-report non-healthy behaviour compared to people that are not trying to lose weight (Lichtman et al., 1992; Muhlheim, Allison, Heshka, & Heymsfield, 1998). Therefore, it is interesting to examine how the effect of the BTS differs for people who are trying to lose weight compared to people who are not. This leads to the second research question:

Does the effect of the Bayesian Truth Serum differ between people who are trying to lose weight and people who are not trying to lose weight?

In order to answer both research questions, a questionnaire study was conducted asking participants about healthy behaviour. Two hundred and twenty-four Dutch female respondents filled out the questionnaire, which consisted of several demographic questions and questions about healthy behaviour. The total sample consisted of females who were trying to lose weight,

the weight-watchers (N = 125) and females who were not trying to lose weight, non-weight watchers (N = 99). For each group the subjects were randomly assigned to either the treatment group (N = 107) or the control group (N = 117). The treatment group was exposed to the BTS as a method to induce more honest and thoughtful responses. However, results indicated that the BTS treatment did not elicit more honest and thoughtful responses to healthy behaviour questions. Furthermore, the effect of the BTS did not differ between weight watchers and non-weight watchers. The presented study was approved by the Ethics Review Board of the Erasmus University Rotterdam.

The paper proceeds as follows. First, in Section 2 relevant literature will be discussed in the theoretical framework. Subsequently, the questionnaire design will be addressed in Section 3. In Section 4 the analyses will be discussed and in Section 5 the obtained results will be presented. The conclusion and discussion will follow in Section 6.

2 Literature Review

This section starts with self-report in healthy behaviour in general, followed by self-reports in physical activity and eating behaviour. Next, the BTS will be explained. Lastly, the hypotheses are developed.

2.1 Self-reports in healthy behaviour

Research examining healthy behaviour tends to rely on self-reports (Hebert et al., 2008; Kristal, Shattuck, & Williams, 1992). Furthermore, self-monitoring methods can be an effective way to lose weight (Burke et al., 2011). Self-monitoring is defined by Foster, Makris, and Bailer (2005) as: “recording dietary intake, and physical activity so that individuals are aware of their current behaviour”. Burke et al. (2011) reviewed 22 studies that examined the effect of self-monitoring on weight loss. From the literature research, Burke et al. (2011) concluded that there is a significant association between self-monitoring and weight loss. With people who self-monitored more frequently, were more successful in losing weight. This suggests that self-monitoring eating behaviour and physical activity can be a useful method to lose weight.

Despite the advantages of self-reports there are also two main problems that occur with self-reports. First, when respondents have to track their habits during a study they might change their behaviour (Macdiarmid & Blundell, 1998). Second, when respondents have to estimate their behaviour they might misreport it. The misreporting can happen conscious or unconscious.

2.1.1 Behavioural change

Respondents could change their behaviour because they become more conscious about their behaviour when they have to self-report it for a study. For example, while self-reporting food intake people are embarrassed by what they usually eat and therefore they will eat less when they are tracking their diet (Macdiarmid & Blundell, 1997). Subjects could also change their behaviour due to convenience. Macdiarmid and Blundell reported this, subjects were more likely to eat the type of foods that were easier to record. Both are reasons for under-eating during the experiment. The change in behaviour is a problem when respondents are asked to report their behaviour beforehand and not when the subjects are asked about their past behaviour. The current study did eliminate the possibility of behavioural change by asking respondents about their behaviour of the last two weeks.

2.1.2 Misreporting

Misreporting is of a concern when behaviour is asked beforehand and when it is asked afterwards. Several factors influence the likelihood someone is misreporting healthy behaviour; these include gender, body mass index (BMI), the desire to lose weight, social desirability bias, and failing to memorise past behaviour correctly.

First, females are more likely to under-report food intake compared to men (Asbeck et al., 2002; de Vries et al., 1994; Haraldsdóttir & Sandström, 1994; Macdiarmid & Blundell, 1998; Novotny et al., 2003). Females are more concerned about their weight and are more affected by social pressure, this is probably influencing the misreporting (Chaiken & Pliner, 1987). Williamson et al. (1992) examined the effect of gender on self-reported behaviour. The study was conducted in the United States, with findings indicating that 25 percent of the males in the sample were trying to lose weight compared to 39 percent of the females. In a more recent study conducted in the United States, Tsai, Lv, Xiao, and Ma (2016) found that males who were overweight were less likely to state that they actively wanted to lose weight (43.3%) compared to overweight females (58.3%). Bunt and colleagues (2017) also investigated the effect of gender on self-reporting behaviour, utilising a Dutch sample. Their findings indicated that females were more likely to say that they would like to have help losing weight (31.5%), compared to males (19.0%). Together, this suggests that females report being more open to getting help losing weight and also more actively want to lose weight.

The second factor is body weight, people who are overweight (BMI $\geq$ 25) or obese (BMI $\geq$ 30) are more likely to misreport healthy behaviour (Prentice et al., 1986; Lichtman et al., 1992; Johnson et al., 1994). The third factor that influences misreporting is whether someone is trying to lose weight or not. Lichtman et al. (1992) studied the effect of under-reporting of food intake and physical activity for diet-resistant obese. These are obese people that stated that they wanted to lose weight but were not able to and they reported a food intake that should have led to weight loss (Muhlheim, Allison, Heshka, & Heymsfield, 1998). However, they did not lose weight because these diet-resistant obese under-reported their food intake by 50 percent and over-reported their physical activity by 50 percent. Another study examined the under-reporting of food intake for non-diet-resistant obese. It was found that the under-reporting of food intake is slightly lower, being 20 to 40 percent, for the non-diet resistant obese (Bandini, Schoeller, Dyr, & Dietz, 1990; Prentice et al., 1986). Under-reporting for non-obese is even lower (Black et al., 1993). It was found that non-obese under-reported food-intake by 0 to 20 percent. Together, these studies indicate that people who are trying to lose weight are more likely to under-report healthy behaviour.

The fourth factor to decreases the trustworthiness of self-reporting is the social desirability bias (Hebert, Clemow, Pbert, Ockene, & Ockene, 1995). When a respondent suffers from this bias she is more likely to report behaviour that is socially admirable. In the perspective of healthy behaviour, the social desirability bias leads to more people self-reporting less food intake and higher amounts of physical activity. The fifth factor is related to cognitive capabilities. Misreporting can occur when subjects fail to correctly memorise their past behaviour (Macdiarmid & Blundell, 1998). In this situation respondents could have the intention to report behaviour accurately but might fail in remembering their food intake or physical activity.

Taken together, these are five common reasons for misreporting healthy behaviour, which will lead to unreliable measures of healthy behaviour. It is therefore essential to examine more deeply how misreporting occurs for physical activity and eating behaviour specifically.

2.2 Over-reporting of Physical Activity

Increasing physical activity is one way to lose weight (Klesges et al., 1992). Physical activity is defined by Shephard (2003) as: “all types of muscular activity that increase energy expenditure substantially” (p. 197). Physical activity can be measured directly by the use of

wearable monitors or indirectly by the use of self-reports (Ainsworth et al., 2015). Which method is prefered depends on the purpose of the study. For measuring energy expenditure, the double-labeled water method is currently the golden standard for free-living respondents (Ainswort et al., 2015). The double-labeled water method is a method to measure energy expenditure by the use of an accelerometer (Schoeller & Van Santen, 1982). Subjects drink water which contains isotopes and then follow their usual habits. Afterwards, the produced CO_2 can be used to calculate energy expenditure. Other methods include the use of pedometers and heart rate monitors (Ainsworth et al., 2015). These methods are relatively more expensive and complicated compared to self-reports.

However, the development of movement tracking applications for smartphones and wearable devices might replace these measures soon. Alharbi et al. (2016) studied the accuracy of the Fitbit. They found that this new measure is accurate, although step counts are slightly over-estimated. Another study tested different smartphones and wearable devices on their accuracy of step count (Case et al., 2015). Subjects had to walk on a treadmill wearing the devices. From this study it was found that both smartphones and wearable devices are close to real step count. However, the smartphones were more accurate. At the moment, smartphones and wearable devices are not accurate enough, but in the future these might be a cost-effective way to measure physical activity. Currently, self-reports are still essential because self-reporting methods offer an inexpensive option that can be used in studies where large population size is of interest (Laporte et al., 1985; Strath et al., 2013). However, a limitation of self-reports is that these are less reliable due to misreporting.

Direct tools to measure physical activity are widely used to validate the reliability of self-reported measures. Gorzelits et al. (2018) examined the discordance between self-reported physical activity and physical activity measured by devices for males and females. On average, 77 percent of the male subjects and 72 percent of the female subjects self-reported to meet the aerobic activity guidelines whereas only 21 percent of the males and 17 percent of the females met the guidelines based on the measurement by the device. From the comparison between direct tools and self-reports, it was found that physical activity was over-reported. The results of the study illustrate the over-reporting of physical activity, the next paragraph discusses several other examples of over-reporting of physical activity.

Rzewnicki, Auweele, and De Boudeaudhuij, (2003) demonstrated that respondents in the International Physical Activity Questionnaire (IPAQ) over-reported physical activity when

compared to respondents in the same IPAQ when interviewers used a probe protocol and probing. In this experiment telephonic interviewing was used, and it was demonstrated that how the interviewers asked questions affected the reported physical activity. In addition, Sims, Smith, Duffy, and Hilton (1999) studied the effect of motivating the elderly to perform more physical activity. In the study, it was found that people reported more physical activity after being motivated. However, the results of the heart-rate data did not confirm this increase. This means that people over-reported physical activity even more when they were motivated to exercise more. The over-reporting could be due to social desirability bias that could have occurred because of a third party that was motivating the elderly. The effect of the social desirability bias was examined by Adams et al. (2005) who studied its effect on over-reporting. Social desirability was measured using the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960). In their study, three types of methods for self-reporting physical activity were compared to an objective measure of physical activity, known as the double-labeled water method. The self-reported measures were a 24-hour physical activity recall, and two types of 7-day physical activity recalls. Only the former was correlated with the double-labeled water method. Furthermore, a regression analysis was conducted to estimate the effect of social desirability bias on self-reported physical activity. This analysis showed that social desirability let to a significant overestimation of physical activity in the 7-day physical activity recall methods. Similar results were found in a meta-analysis comparing 31 studies in which a social desirability scale was used (Van de Mortel, 2008). Results showed that in 43% of the studies the social desirability bias affected the results. This means that social desirability plays a significant role in over-reporting of physical activity in self-report studies. Currently, objective measures of physical activity should check if a questionnaire gives reliable results or not.

2.3 Under-reporting of Eating Behaviour

For eating behaviour, the use of objective measures to monitor behaviour is also complex. The double-labeled water method can be used (Jinnie, 2015) or accelerometers (Samuel-Hodge et al., 2004). However, these measures are expensive and therefore not effective for large samples. The development of an alternative measure for eating behaviour is promising. Recently, the Remote Food Photography Method (RFPM) was developed (Martin et al., 2012). Respondents were asked to take pictures of their food and the leftovers, and send these to a server. The server would calculate food intake almost real-time. Martin and colleagues did a pilot study to test this method by comparing it to the double-labeled water method. At first, a significantly different effect was found, indicating that the RFPM was not

reliable. After altering the method the results of a second study gave reliable results. In another study the RFPM was validated for adults and children (Martin et al., 2014). Here it was found that the RFPM is a reliable measure. All in all, this new method is a promising alternative for expensive measures as the double-labeled water method. However, at the moment self-reported measures are still the easiest to conduct for a large sample.

Therefore, a reliable self-report would be the most straightforward measure of eating behaviour. This is important since someones' eating behaviour influences his or her weight (Klesges et al., 1992). However, previous research has demonstrated that self-reports do not always accurately reflect eating behaviour due to participants often under-reporting their food intake (Johansson et al., 2001). Under-reporting of energy intake takes place within different groups within a given population but tends to differ by personal characteristics.

Hill and Davies (2001) examined the underlying personal characteristics that led to under-reporting. They argue that the under-reporting of food intake might not be random. Specifically, the higher someone's BMI, the more likely someone is to under-report (Heitmann & Lissner, 1995). More recently, Freisling et al. (2012) found the same result in a study across six European countries. These findings also seem to apply to the Netherlands, where Trijsburg et al. (2017) also found that Dutch people with a higher BMI are more likely to under-report their food intake. Collectively, these studies suggest that self-reported measures might be biased, and these biases might be stronger for people with a higher BMI. As a result, it is especially problematic to rely on self-reports when studying healthy behaviours in overweight people. Reducing these biases could potentially improve self-monitoring methods.

Another factor that influences under-reporting of food intake is the social desirability bias (Hebert, Clemow, Pbert, Ockene, and Ockene, 1995). This potential bias that can occur with self-reported behaviour was confirmed by Schoeller (1990). He studied how people report their consumption. In this study, he found that self-reported dietary intake is closer to the consumption levels that subjects perceive as socially acceptable rather than to their actual dietary intake levels. The reported dietary intake was under-reported for the nine examined studies by Schoeller. In addition, females are more affected by social desirability bias compared to men (Hebert et al., 1995). Therefore, in this study, we chose to have a complete female sample, since the effect of the BTS would be higher. All in all, the self-reports should be improved by reducing the factors that influence under-reporting of food intake. This paper examines whether the BTS is an effective method to accomplish this.

2.4 Bayesian Truth Serum

From the above paragraph, it was made clear that there is a need to increase the reliability of self-reports about healthy behaviour. The current section explains a method which was developed to increase truth-telling behaviour in self-reports. In an effort to provide a measure for subjective truth, Prelec (2004) invented the Bayesian Truth Serum (BTS). The BTS is a method that elicits truthful answers when the objective truth is hard or impossible to know. This method is based on the assumption that people believe that their truthful answer is more common than collectively predicted. This section continues as follows. First, the application of the BTS in other fields is discussed. Next, the necessary information about the BTS is given, followed by the discussion of the 'false' consensus effect. Furthermore, the BTS-formula is explained, followed by the underlying assumptions.

2.4.1 Application in Other Fields

The BTS-method has already been applied to several fields, including contingent valuation, deterrence studies, new product forecasts, recognition questionnaires, and unethical research practices (Barrage, & Lee, 2010; Howie, Wang & Tsa, 2011; Loughran, Paternoster, & Thomas, 2014; John, Loewestein, & Prelec, 2012; Weaver & Prelec, 2013).

Barrage and Lee (2010) applied the BTS in the field of contingent valuation. A problem in studying contingent valuation is the hypothetical bias. This is the bias that respondents might answer differently when no real monetary incentive is used (Cummings et al., 1997). It was tested whether the use of the BTS could eliminate the hypothetical bias. In the study by Barrage and Lee (2010), they let respondents vote about the division of money over two charities. Five treatments were used, a control treatment, a hypothetical treatment, cheap talk, consequential treatment, and the treatment of interest the BTS. It was found that the BTS only decreased the hypothetical bias for one of the two charities. However, the BTS had a significant effect on females.

Another field in which the truth is essential is the field of criminology. Loughran, Paternoster, and Thomas (2014) applied the BTS to this field. In their study subjects were randomly assigned to the control group or the BTS group. Subjects in the control group were given an incentive for participating in the experiment, whereas subjects in the BTS group were given an incentive based on whether they answered honestly and thoughtfully. All subjects were asked to self-report willingness to offend. It was found that the BTS affected self-reports in some of the offendings. For example, in the BTS group more people admitted that they would

be willing to drive drunk and cheat on an exam. However, for texting while driving and smoking marijuana the control group and the BTS group did not give different answers. Therefore, it can be concluded that the BTS works for some questions in the field of criminology, however, not for all questions.

Howie, Wang and Tsai (2011) studied whether the BTS could increase the reliability of new product forecasts. In this field it is not possible to know the truth objectively. In their study a survey using the BTS question format was used and compared with a reference model to predict product adoption, and it was also compared to actual adoption of the products asked about during the survey. For the BTS prediction subjects where weighted based on their BTS score, this improved the forecast of product adoption significantly.

Weaver and Prelec (2013) examined the BTS in a setting were the objective truth could also be partly known. They analysed the truth-telling method in a recognition questionnaire where different items were presented. Among these items were foils; brands or scientific terms that do not exist. Respondents that were exposed to the BTS recognised fewer foils compared to the control group. Another study using the BTS-method was conducted by John, Loewestein, and Prelec (2012). In their study, they surveyed psychologists about unethical research practices. The group that was exposed to the BTS admitted to the unethical research practices more than the control group. When the sensitive questions were asked without the BTS, people answered less truthfully.

For all these findings it was hard or even not possible to know the truth objectively. The findings discussed here confirm that the BTS can elicit more truthful responses. It is essential to obtain the truth in questionnaires in order to have reliable research results. Therefore, it is of interest to examine in which fields of research the BTS can also elicit more truthful answers. In this study the BTS will also be implied to healthy behaviour questions in an attempt to elicit more truthful responses.

2.4.2 BTS Basics

In a BTS questionnaire, two types of questions are asked; one question about the subjects' behaviour and one question about what the subject expects to be the behaviour of other subjects. The BTS method calculates an information score based on the answers to the questions about someone's behaviour. This score is based on the estimation of the behaviour of others and the actual reported behaviour. The scoring system rewards truthful answers with higher scores. Subjects could get a monetary incentive based on the BTS score, or they could

influence the amount of money donated to a charity. The latter was used in this study in accordance with John et al. (2012). The preference of subjects with higher scores got a higher weight in deciding on the division of money over three charities. More details about the information scoring can be found in Section 2.5.4 and more details about the incentives can be found in Section 3.3.2.

The next questions illustrate the BTS method. First, a question about individual behaviour was asked, followed by an estimation of the percentage of others that exhibit the same behaviour.

Do you agree with the following statement: "In general, I eat healthy"? (1)

- Agree
- Disagree

Please estimate, how many percent of other respondents in this survey agree with this statement: (2)

How many percent: (1-99%)

Based on these two questions the BTS-score can be calculated. This score consists of an information score and a prediction score. The information score relies on the answer a subject gives (1), and the prediction score relies on the subjects' estimation of the frequencies (2).

2.4.3 The 'False' Consensus Effect

The underlying reasoning behind the BTS method is the 'false' consensus effect. Ross, Greene, and House (1977) define the false consensus effect as: "to see their own behavioural choices and judgments as relatively common and appropriate to existing circumstances while viewing alternative responses as uncommon, deviant, or inappropriate" (p. 280). Hoch (1987) found that people tend to take into account their own behaviour when estimating the behaviour of others. Hoch also looked at the accuracy of the predictions that were made. He found that predictions became more accurate when someone relied on his or her behaviour to predict the behaviour of other, similar, subjects. As a consequence, the false consensus effect does not have to be 'false' according to Dawes (1989). The false consensus effect leads to better predictions in Bayesian Models. He defines the false consensus effect as: "an egoistic bias to overestimate the degree to which others are like us." (p. 1). People who rely on their behaviour in making

predictions about the behaviour of others are implicitly using the consensus effect, and might make better predictions.

How the false consensus effect applies to the BTS is explained using the example questions from Section 2.5.1. Females who agree with the statement "In general, I eat healthy" are expected to estimate a higher percentage of the other respondents in the experiment that also agree with this statement compared to females who do not agree with the statement. The healthy eaters are believed to consider their behaviour when estimating the percentage of other healthy eating respondents. This is referred to as the 'the false consensus effect'. Even if the healthy eaters assume that they are the minority of the population, their estimation would be higher than the average estimation of non-healthy eaters. This is true since an individual will expect that her answer (healthy behaviour) is more common than collectively predicted (Prelec, 2004). This way, telling the truth will maximise the BTS score of an individual, and therefore will be the best strategy since the money that will be divided over the charities is more likely to go to the charity of the subjects' choice.

2.4.4 BTS-Formula

The BTS-Formula is only maximized when respondents answer truthfully. This BTS scoring method was proposed by Prelec (2004). Respondents are denoted by $r \in \{1, 2, \dots\}$. First, respondents answer a question about their own behaviour, which results in a truthful answer to a two-multiple-question, denoted by $t^r = \{t_1^r, t_2^r\}$. The variable x_k^r gives an indication of the answer-option that is selected, being 1 when the answer is given and 0 otherwise. Next, respondents give a prediction of the sample population $(y_1^r, \dots, y_m^r)$. When the common prior assumption holds, respondents with the same truthful answer will estimate the same population distribution. When $t^r = t^s$ then also $p(\omega|t^r) = p(\omega|t^s)$ most be true. Therefore, respondents can be generalised into two categories, people who agree with a statement and people who do not.

From the data generated the predicted and actual estimation of the total population can be calculated. The actual percentage of people who admit to a particular behaviour proposed by answer-option k is denoted by $\bar{x}_k$. The geometric average of the predicted percentage of people who would admit to a certain behaviour proposed by answer-option k is denoted by $\bar{y}_k$. The information score (I) is then given by the log-ratio of actual-to-predicted frequencies. So, $log\ \frac{\bar{x}_k}{\bar{y}_k}$ is the information score for answer *k*. The respondent is also rewarded for her prediction;

this is indicated by the prediction score (P). The prediction score rewards accurate predictions by giving higher scores to respondents who better predict the average percentage of respondents who exhibit a particular type of behaviour. The BTS score for a respondent can be calculated by adding the information score (I) to the prediction score (P), which gives the following formula:

$$score\ for\ respondent\ r = \sum_k x_k^r log \frac{\bar{x}_k}{\bar{y}_k} + \alpha \sum_k \bar{x}_k log \frac{y_k^r}{\bar{x}_k}, 0 < \alpha$$

In this study alpha (α) is assumed to be equal to one since no prior information was found about the study of the BTS in the healthy behaviour setting. By assuming $\alpha = 1$ the information score and prediction score are given an equal weight.

In order to illustrate how the BTS-formula works, an example of the information score and an example of the prediction score is given below.

The information score (I) of an individual depends on the answer she gives (x_k^r), the actual endorsed behaviour of the population $(\bar{x}_k)$, and the geometric average of the estimated behaviour of the population $(\bar{y}_k)$.

$$I = \sum_k x_k^r \log \frac{\bar{x}_k}{\bar{y}_k}$$

For example, when the geometric average of estimated behaviour is equal to 25 percent and the actual endorsed behaviour is 30 percent. For an individual, endorsing the same behaviour, the information score will be:

$$I = \sum_k 1 \cdot \log \frac{30}{25} = 0.079$$

In this case, the actual endorsed behaviour is more common than collectively predicted. Therefore the information score is high. Instead, when the actual endorsed behaviour is less common than collectively predicted, the information score would be low. This would be the case when the collectively predicted frequency would be 50 percent:

$$I = \sum_k 1 \cdot \log \frac{30}{50} = -0.222$$

The prediction score (P) rewards predictions of the endorsed behaviour by the population that are closer to the actual frequencies. It depends on the prediction of an individual respondent (y_k^r) and the actual endorsed behaviour frequency ($\bar{x}_k$).

$$P = \sum_k \bar{x}_k log \frac{y_k^r}{\bar{x}_k}$$

So, when the actual frequency would be equal to 30 percent, and someone estimates it to be 30 percent, the prediction score is at its maximum, equal to zero:

$$P = \sum_k 30 \cdot log \frac{30}{30} = 0$$

However, when someone predicts the frequency to be 50 percent, the information score will be lower:

$$P = \left(30 \cdot log \frac{50}{30}\right) + \left(70 \cdot log \frac{50}{70}\right) = -3.574$$

In summary, an answer gets a higher information score when it is surprisingly common and the closer the predicted frequency is to the actual frequency, the higher the prediction score (Prelec, 2004).

2.4.5 BTS Assumptions

The BTS- model relies on several assumptions that will be addressed in this section. First, the information-scoring method relies on the assumption that the population consists of rational individuals who maximise their expected outcome (Prelec, 2004). Second, the sample should be large enough so that one single answer cannot noticeably influence the empirical frequencies. Third, the common prior assumption must hold. Meaning that common knowledge is assumed and respondents will believe that other respondents that share their opinion will come up with the same estimations of the entire population. Therefore, respondents use their opinion as an "impersonally informative" signal about the behaviour of other respondents in the sample. This assumption can be tested by comparing the average prediction for respondents who exhibit the same behaviour and respondents who do not exhibit the same behaviour. When these three assumptions hold, giving truthful answers would be the optimal strategy when someone believes that other respondents also answer truthfully.

2.5 Hypotheses

For both physical activity and eating behaviour the objective truth is knowable when expensive and time-consuming measures are used. However, for studies concerning larger populations with a smaller budget this is a problem. Self-reports are in this situation easier and cheaper to conduct but lead to a subjective truth that may not reflect reality. Since the Bayesian Truth Serum (BTS) rewards truthful answers in an objective way, this method can be used to elicit more truthful answers about healthy behaviour (Prelec, 2004). Although the name of the BTS might suggest that it is a method to let liars tell the truth, it is more refined. The BTS pushes people to really think about their answers to each question and it attempts to prevent biases. Therefore, the BTS method will be used in an attempt to decrease over-reporting of physical activity and under-reporting of eating behaviour. Earlier studies have shown that the BTS can elicit more honest answers to non-desirable behaviour questions, even about sensitive topics such as unethical research practices in academia (John et al., 2012). Since non-healthy behaviour can be seen as socially undesirable behaviour, it is likely that the BTS leads to a higher number of respondents admitting to non-healthy behaviour, leading to the following hypothesis:

H_1: *Respondents in the BTS condition are more likely to report non-healthy behaviour compared to respondents in the control condition*

The difference between the non-incentivised self-report of healthy behaviour and the BTS-incentivised self-report of healthy behaviour might be greater for people that are trying to lose weight compared to people who are not trying to lose weight (Freisling et al., 2012; Heitmann & Lissner, 1995; Trijsburg et al., 2012). Lichtman et al. (1992) suggest that diet-resistant people tend to under-report food intake most. In other words, people who are trying to lose weight could be more likely to report desirable behaviour and therefore misreport the most. Therefore, the following hypothesis is proposed:

H_2: The difference in responses to healthy behaviour questions between the BTS and the control group is higher for weight watcher than for non-weight watchers

3 Experimental Design

This experiment seeks to measure the effect of the BTS on self-reported healthy behaviour. A 2 (BTS - Control) by 2 (weight watcher – non-weight watcher) between subject design was used. To this end, a questionnaire was constructed with two treatment conditions (i.e., BTS-treatment and control condition). Subjects were categorised into weight watcher or non-weight watcher depending on how they answered a question about their behaviour. Afterwards, subjects in each group (weight watcher or not) were randomly allocated to either the BTS group or the control group.

3.1 Materials

The full questionnaire constructed to investigate the research questions can be found in Appendix A. The questionnaire started with an introduction, followed by exclusion criteria (i.e., Dutch, female, and no eating disorder). Subjects who did not match these criteria were not able to take part. Next, demographic questions were asked including age, education, weight, and height. Afterwards, a question regarding the current weight loss status of a respondent was asked. Next, for each group, weight watchers and non-weight watchers, respondents were randomly divided between the BTS group and control group. The division was done separately in order to have an equal division of respondents who were in the BTS treatment and in the control group, for both the weight watchers and the non-weight watchers. After the division, subjects got the questions about healthy behaviour and for the BTS group these were followed by the BTS questions.

3.1.1 Current weight loss status

To identify which participants were actively trying to lose weight, four statements were presented. These are based on the statements previously used by Kuijer and Boce (2014). Respondents were asked to indicate with which statement they identified most: “I am trying to lose weight”, “I am trying to maintain a healthy weight”, “I am trying to gain weight”, and “I am not watching my weight”. Respondents who selected either the first or the second option were classified as ‘weight watchers’, whereas respondents who chose the third or fourth option were classified as ‘non-weight watchers’. These two different groups were used to compare the effect of the BTS for subjects who are watching or not watching their weight. Randomisation between the control and BTS-treatment was implemented randomly for the two groups separately. This way, it was ensured that in each group there would be about the same amount of respondents.

3.1.2 Control and BTS instructions

In line with John et al. (2012), participants were incentivised by donation money (€50) on their behalf to charity. In the control condition, respondents could not influence how much money was given to each charity. However, in the BTS condition, participants could influence which charity would receive the money. By answering the questions honestly, they received a higher BTS score, and their answers would be weighted higher than the answers of respondents with lower BTS scores. In summary, respondents in the control group had an incentive to participate and subjects in the BTS group got an incentive to answer truthfully. The BTS instructions are as well based on the paper by John et al. (2012), see below:

"Let us apply a formula called the Bayesian Truth Serum," which would be used "to determine the size of the donation made to the charity that you selected." They were also told that "the important property of the formula is that it rewards truthful answers. This means that truthful answers about your practices will increase the donation made on your behalf (and will also tend to increase the donations made on behalf of other respondents). For the purpose of this survey it is not necessary for you to understand how the formula works, although the theoretical paper from Science, which includes a short abstract, is available here (link to paper)." (John et al., 2012)

In line with the questionnaire of John et al. (2012) a check-question was inserted to ensure participants understood the explanation of the BTS. It is essential to know whether subjects understand the basic idea of the BTS in order to let them answer more honestly. Therefore, as proposed in the questionnaire of John and colleagues this research also used a question to check this. Subjects were asked to fill out the following statement: "Giving truthful responses on this survey _____ the amount of money donated to charity on my behalf. Choosing from the response options: *has no impact on, decreases, increases*." (John et al., 2012, p.4). If respondents did not answer the question correctly, they were redirected to the instruction page. When a respondent answered correctly, she went to the next part of the questionnaire with the healthy behaviour, and BTS questions. In the rare occasion that a respondent answered the question wrong twice ($N = 2$), she was redirected to the end of the survey and her response was not recorded.

3.1.3 Healthy Behaviour questions

After the instructions about the incentives, the healthy behaviour questions were asked. The first question was about the general eating behaviour of a subject. The question from Van

Beek, Antonides, and Handgraaf. (2013) was used: “Do you agree with the following statement: In general, I eat healthy?” (p. 786). Van Beek et al. (2013) used recommendations from the Dutch Nutrition Centre for their other questions about healthy eating behaviour. These same questions were used in this research. For example: “Please, indicate how many days of the week you consumed at least 200 grams of fruit.” The question about vegetable consumption was slightly altered. The Dutch Nutrition Centre (2016) have since adjusted their recommended daily dose of at least 200 grams of vegetables to 250 grams. Therefore, this new recommendation was used when creating the questionnaire for this research. The full list of questions is shown in Appendix A.6. Participants were asked on how many days of the week they match these consumption criteria. After the data was collected, the answers were converted to binary variables, indicating ‘yes’ when the criteria are met, and ‘no’ when they are not met. Dassen, Houben, and Jansen (2015) utilized the questions created by Van Beek et al. (2013), also making minor alterations to the questionnaire for their research.

The physical activity part of the healthy behaviour questions started with the following question. “Do you agree with the following statement: In general, my physical activity is sufficient?” (Van Beek et al., 2013, p. 786). The answer to this question shows the self-reported general physical activity of a respondent. Next, two questions based on recommendations were asked. In the Netherlands, it is recommended by the Health Council of the Netherlands (2017a) to be moderately physically active for at least 150 minutes a week. Another recommendation is to perform intensive physical activity for 20 minutes, at least two times a week. These recommendations were based on literature research by the Health Council of the Netherlands (2017b). Van Beek et al. (2013) used recommendations from the Dutch Institute for Exercise and Physical Activity in their questionnaire to estimate someone’s physical activity. In line with Van Beek et al. (2013), these type of questions were asked. For example, subjects were asked to indicate on how many days a week they were moderately physically active for at least 30 minutes. See Appendix A.6 for the complete list of questions.

3.1.4 BTS-questions

In the control condition, participants were only asked about their own behavior. In the BTS treatment, after each healthy behaviour question, an additional BTS-question was asked. These were based on the questions asked by John et al. (2012). It was asked what percentages of the other respondents would have exhibited a specific behaviour. For example: “Please estimate, how many percents of other respondents in this survey agree with the statement “In general, my physical activity is sufficient.”. The answer format was a scale ranging from 1 to

99 percent, this scale is used for mathematical purposes. An estimation of 0 or 100 percent would mean that there is no deviation in opinions, this is highly unlikely. Therefore, the scale could be ranged from 1 to 99 percent. The questions can be found in Appendix A.7.

3.2 Participants

An online survey was used as a means of collecting the data for this research. The survey was hosted on Qualtrics from June 2nd to June 9th. Respondents were recruited via Facebook. To increase the power of the statistical tests that were conducted, blocking was used. Blocking is selecting participants with particular characteristics in such a way that a homogeneous sample is created. By targeting a specific group of subjects the differences between the groups were more likely to come from the treatment than from individual differences. For gender, only females were chosen since females are more willing to seek help when trying to lose weight (Bunt et al., 2017). Furthermore, to only include respondents who are capable of having healthy eating and exercising behaviour, respondents were asked whether they have an eating disorder such as Anorexia Nervosa or Bulimia Nervosa. Respondents who are male, or have Anorexia Nervosa or Bulimia Nervosa were immediately redirected to the end of the survey.

In total, 244 females started the questionnaire. The total drop-out rate of the experiment is 8.20%. From the 8.20%, 4.51% of the drop-outs happened before the participants were randomly allocated into either the BTS or the control group. After the division, the drop-out rate was 7.76% in the BTS group and 0% in the control group. This drop-out rate includes two subjects who were removed for the experiment during the questionnaire because they did not answer the BTS check question correctly. The final sample consisted of 107 respondents in the BTS group and 117 respondents in the control group (Table 1).

	Frequency	**Percent**	**Cum.**
Control	117	52.2	52.2
BTS	107	47.8	100.0
Total	224	100	

Table 1: *Descriptive Statistics for BTS*

Several variables were derived from the questions in the survey, namely height, weight, age, education, and current weight loss attempt (Tables 2-4). BMI is a continuous variable calculated by weight and height using the following formula: $\frac{(weight\ in\ kg)^2}{height\ in\ cm \cdot height\ in\ cm}$. Age is a continuous variable giving the age of a participant in years. Education is a categorical variable indicating the level of education someone has obtained. Current weight loss attempt is measured

over four categories which have been merged into two groups: weight watchers (group 1 and 2), and non-weight watchers (3 and 4).

Variable	Obs.	Mean	Std. Deviation	Min	Max
BMI	224	24.049	0.733	15.57	43.23
Age	224	30.99	13.765	17	64

Table 2: *Descriptive Statistics for BMI and Age*

	Frequency	Percent	Cum
Actively trying to lose weight (1)	65	29.0	29.0
Actively trying to maintain current weight (2)	60	268	55.8
Actively trying to gain weight (3)	7	3.1	58.9
Not actively watching weight (4)	92	41.1	100.0
Total	224	100	

Table 3: *Descriptive Statistics for current weight loss attempt*

	Frequency	Percent	Cum
BTS & Weight watcher (1)	60	26.8	26.8
BTS & Non-weight watcher (2)	47	21.0	47.8
Control & Weight watcher (3)	65	29.0	76.8
Control & Non-weight watcher (4)	52	23.2	100.0
Total	224	100	

Table 4: *Descriptive Statistics for BTS_WeightWatcher*

The mean age of the four groups does not significantly differ (Table 5). However, the results from the One-way ANOVA indicate that the average BMI differs among groups (p = 0.018). In both weight-watchers groups the BMI is higher ($M_1 = 25.276, M_3 = 24.390$) compared to the non-weight watchers groups ($M_2 = 22.693), M_4 = 23.431$). The difference between the group is as expected because the division into weight-watcher and non-weight watcher depended on whether someone wants to maintain her current weight or not. Respondents with a higher BMI are more likely to be watching their weight then respondents with a lower BMI. Therefore, the difference between these groups is justified since we are interested in this difference and the effect that the BTS has on it. For education the frequencies between the groups do not differ significantly (Table 6). In summary, the four groups do not differ based on age and education. The groups differ based on BMI but this is essential to answer the second research question.

	Age	BMI
BTS & Weight watcher (1)	31.48	25.276
BTS & Non weight watcher (2)	30.28	22.693
Control & Weight watcher (3)	32.77	24.390
Control & Non-weight watcher (4)	28.85	23.431
Sig.	0.469	0.018*

Table 5*: Results of the One-way ANOVA for age and BMI, given means of age and BMI*

Education frequency	BTS & Weight watcher	BTS & Non-weight watcher	Control & Weight watcher	Control & Non-weight watcher	Total sample
Vmbo, mbo 1, avo, onderbouw	2	2	1	3	8
Havo, vwo, mbo	21	17	29	24	91
Hbo-bachelor, wo-bachelor	29	21	27	22	99
Hbo-master, wo-master, doctor	8	7	8	3	26
Total	60	47	65	52	224
Sig.	0.799				

Table 6: *Test results of the Pearson Chi-Square test for the frequencies for educational attainment*

4 Analyses

To test whether the BTS could elicit more reliable responses in healthy behaviour questions several analyses were conducted. For all hypothesis testing, the Bonferroni Correction was used. By making multiple comparisons using the same sample, the likelihood of one of these comparisons yielding a significant effect increases. Therefore, an adjustment was made to the critical alpha to make sure that the chances of obtaining type I errors were reduced. The critical alpha is adjusted to 0.008 (0.05/6) since six comparisons were made.

In order to check the common prior assumption, the t-test or Mann-Whitney U test will be used, depending on whether the assumptions for a parametric test are satisfied. When the results of these tests show that there are significant differences at the five percent significance level, then it can be concluded that the common prior assumption holds.

To identify whether respondent in the BTS group self-reported more unhealthy behaviour, compared to respondents in the control group, several tests were conducted. First, six Pearson Chi-square tests were performed with BTS as the independent variable and the six binary healthy behaviour types (general eating behaviour, fruit consumption, vegetable consumption, general physical activity, intensive physical activity, and intensive physical activity) as dependent variables. Next, to ascertain the effect of BTS, age, BMI and education on the six healthy behaviour types, six binary logistic regressions were performed. Weaver and Prelec (2013) also used binary logistic regressions in order to examine the effect of the BTS.

In order to examine whether the effect of the BTS is higher for weight watchers compared to non-weight watchers, six log linear models were conducted. These examined the relationship between the BTS, weight watching, and each of the healthy behaviour types. In addition, six binary logistic regressions were performed to include the effect of the demographics as control variables.

5 Results

The results section starts by evaluating the BTS assumptions, in particular the common prior assumption was tested. Next the results from the tests about the first hypothesis are discussed, followed by the results of the second hypothesis.

5.1 BTS-assumptions

The first assumption for the BTS model is that the sample consists of rational individuals. This assumption could not be tested. The second assumption for the BTS is that the sample should be large enough. The current sample consists of 224 respondents. Compared to John et al. (2012) this is a rather small sample because they used a sample of 2,155 respondents. Therefore, it cannot be concluded that this assumption holds.

The third assumption is the common prior assumption, which was tested. Whether the t-test or the Mann-Whitney U test had to be used was estimated by checking the assumptions of the t-test. The use of the t-tests comes with several assumptions. The estimated percentage is measured on a continuous scale, the independent variable is whether someone shows healthy or unhealthy behaviour, respondents can only show one type of behaviour. Furthermore, the Shapiro-Wilk test (Appendix B, Table 1) was performed to test whether the estimated percentage of healthy behaviour is normally distributed for each group in the healthy/unhealthy category. From the Shapiro-Wilk test it is known that only the estimated percentage of general eating behaviour is not normally distributed. Therefore, the non-parametric equivalent of the t-test, the Mann-Withney U test, should be used for this behavioural type. To test the homogeneity of variance assumption the Levene's test was used (Appendix B, Table 2). The results of these tests indicate that this assumption holds. In conclusion, the common prior assumption had to be tested with a Mann-Withney U test for general eating behaviour (Appendix B, Table 3), and with five independent t-tests for the other type of behaviours (Appendix B, Tables 4-8).

When the common prior assumption holds, females who exhibit a particular behaviour would also predict a higher percentage of other females to exhibit this same behaviour in comparison to females who exhibit another type of behaviour. For example, when someone eats healthy, she would predict a higher percentage of other respondents in the sample to eat healthy in comparison to someone who does not eat healthy. This consumption was tested by one Mann-Withney U test and five independent t-tests, the results are graphically illustrated in Figure 1.

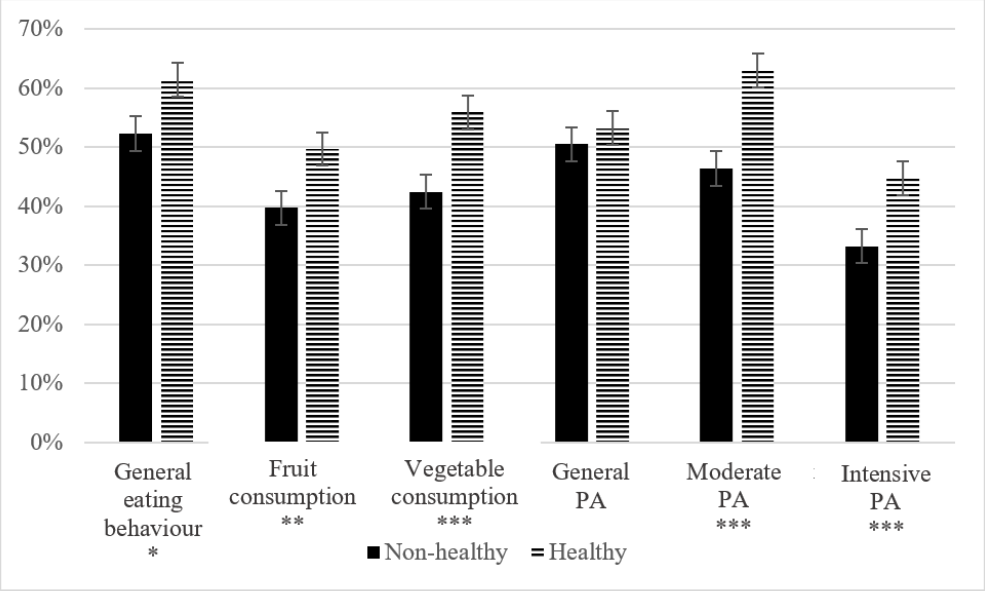

Figure 1: *The average estimated percentage of healthy behaviour by self-reported behaviour* *** *indicates a Bonferroni significance level of 0.01/6 = 0.0017,* ** *indicates a Bonferroni significance level of 0.05/6= 0.008,* * *indicates a Bonferroni significance level of 0.10/6=0.017 for the t-test*

Figure 1 illustrates the result that females who self-reported healthy behaviour would also estimate a higher percentage of other respondents to behave healthily compared to females who self-reported non-healthy behaviour. This effect was significant for eating behaviour (Z = -2.393, p = 0.017, $\eta^2 = 0.054$), fruit consumption ($t(105)$ = -2.777, p = 0.007, d = 0.536), vegetable consumption ($t(105)$ = -3.619, p = 0.000, d = 0.698), moderate physical activity ($t(105)$ = -4.646, p = 0.000, d = 0.907), and intensive physical activity ($t(105)$ = -3.456, p = 0.000, d =0.684). However, for general physical activity no significant effect was found ($t(105)$ = -0.659, p = 0.511, d = 0.171). The t-test for general physical activity indicates that females who self-report healthy intensive physical activity, do not estimate a significantly higher percentage of other females exhibiting the same behaviour (M = 53.276), compared to females who self-report unhealthy intensive physical activity (M = 50.500). Taken together, these results suggest that the common-prior assumption holds for the behavioural questions about general eating behaviour, fruit consumption, vegetable consumption, moderate physical activity, and intensive physical activity. However, the common-prior assumption does not hold for the question about general physical activity. Therefore, the effect of the BTS for the general physical activity question needs to be interpreted with caution.

In summary, the first and third assumption might hold but cannot be tested. The second assumption holds for five out of the six questions. The common prior assumption does not hold for the general physical activity question.

5.2 Difference between BTS and control

To test whether respondents in the BTS condition are more likely to report non-healthy behaviour compared to respondents in the control condition, six Pearson Chi-Square tests were performed. It was tested whether being in the BTS or control group significantly related to the self-reported behaviour, which could be either healthy or non-healthy. Respondents could only belong to one of the four groups, and the variables are measured on a nominal scale. Therefore, the assumptions for a Pearson Chi-Square tests hold. The difference between the BTS group and treatment group is graphically illustrated in Figure 2, the results of the six Pearson Chi-Square tests can be found in Appendix C Tables 1-6.

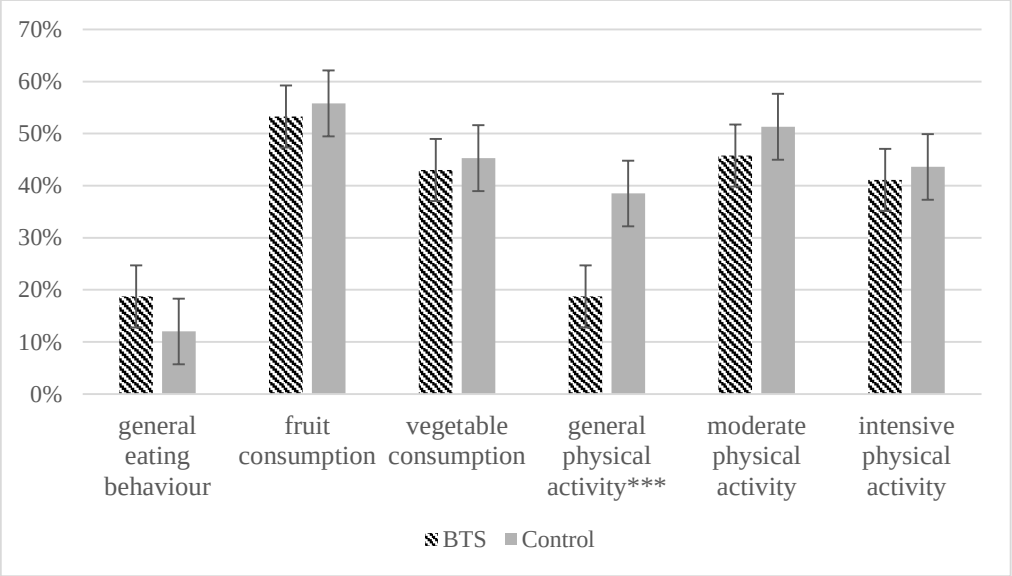

Figure 2: *Percentage of respondents admitting to non-healthy behaviour,* *** *indicates a Bonferroni significance level of 0.01/6 = 0.0017 for the Chi-Square test*

The results of the Pearson Chi-Square tests indicated that there is no statistically significant association between whether someone is in the BTS or control group on self-reporting healthy eating behaviour ($\chi^2(1) = 1.964, p = 0.161, \varphi_V = 0.094$), self-reported fruit consumption ($\chi^2(1) = 1.564,\ p = 0.211, \varphi_V = 0.084$), and self-reported vegetable consumption ($\chi^2(1) = 0.121, p = 0.728, \varphi_V = 0.023$). In conclusion, no significant effect of the BTS on self-reported eating behaviour was found.

The test results for moderate physical activity and intensive physical activity also do not give significant effects. Specifically, there is no statistically significant association between whether someone is in the BTS or control group on self-reported moderate physical activity ($\chi^2(1) = 0.674, p = 0.412, \varphi_V = 0.055$) and not on self-reported intensive physical activity ($\chi^2(1) = 0.139, p = 0.709, \varphi_V = 0.025$). However, for general physical activity a significant

effect was found. There is a statistically significant relationship between whether someone is in the BTS group or in the control group, and self-reported physical activity ($\chi^2(1) = 10.605, p = 0.001,\ \varphi_V = 0.218$). Particularly, respondents in the BTS group are less likely to report non-healthy behaviour, compared to respondents in the control group.

The results of the different Pearson Chi-Square tests only showed a significant relationship between the BTS and general physical activity. However, the Pearson Chi-Square test only takes the treatment effect on the outcome variable into account. A set of logistic regressions were performed to also take into account the personal characteristics. Before the result of the logistic regression can be interpreted, six Box-Tidwell tests were performed in order to check whether the assumption of a linear relationship between the continuous independent variables and the logit holds. From the six Box-Tidwell tests (Appendix D, Tables 7-12) it can be concluded that the assumption for the logistic regressions holds, as there is a linear relationship between the continuous independent variables age and BMI, and the logit transformation of the behaviour. This can be concluded since none of the interactions showed in the Box-Tidwell tests is significant.

Six logistic regressions were conducted, three for eating behaviour and three for physical activity (Table 7). The first binary logistic regression was performed to ascertain the effect of BTS, age, BMI, and education on the likelihood that a respondent would self-report healthy eating behaviour. The logistic regression model was statistically significant, $\chi^2(6) = 19.028, p = 0.004$. The model explained 14.2% (Nagelkerke R^2) of the variance in self-reported eating behaviour. Furthermore, it correctly classified 85.3% of the cases, which is better than when it would be predicted by tossing a coin. It was found that BMI and age do have an effect on the likelihood that respondents self-report healthy eating. Females with a higher BMI are less likely to self-report healthy eating behaviour (p = 0.007). Furthermore, females who are older are more likely to self-report healthy eating behaviour (p = 0.007). The second logistic regression is the one that estimates the effect of BTS, age, BMI, and education on the likelihood of self-reporting healthy fruit consumption. No statistically significant effect was found, $\chi^2(6) = 5.261, p = 0.511$. The model explained only 3.1% of the variance in self-reporting fruit consumption, and correctly classified 60.3% of the cases. The third logistic regression was created to ascertain the effect of BTS, age, BMI, and education on the likelihood of self-reporting healthy vegetable consumption. The model however, was not statistically significant, $\chi^2(6) = 8.630, p = 0.195$. It only explained 5.1% of the variance in self-reported

healthy vegetable consumption and correctly classified 61.6% of the cases. Taken together, the results suggest that only for general eating behaviour a reliable prediction for the likelihood to report healthy eating behaviour can be made based on BMI and age. The BTS did not have a significant effect on self-reported eating behaviour.

The last three logistic regressions contain the results of the physical activity questions. The fourth logistic regression of BTS, age, BMI, and education on the likelihood to report healthy general physical activity is statistically significant, $\chi^2(6) = 16.701, p = 0.010$. The model explained 10.3% of the variance in self-reported general physical activity and correctly classified 70.1% of the cases. Respondents in the BTS group are more likely to report healthy physical activity, compared to the females in the control group. The fifth logistic regression examined the effect of BTS, age, BMI, and education on the likelihood to self-report healthy moderate physical activity. The obtained effect was not statistically significant, $\chi^2(6) = 7.334, p = 0.291$. The model explained 4.3% of the variance in self-reported moderate physical activity, and classified 58.9% of the cases correctly. The sixth binary logistic regression estimated the effect of BTS, age, BMI, and education on the likelihood to report healthy intensive physical activity. The founded effect was not statistically significant, $\chi^2(6) = 9.397, p = 0.152$. The model explained 5.5% of the variance in reported physical activity, and correctly classified 57.1% of the cases. The results of the physical activity questions indicate that only the general physical activity question could be predicted by the BTS. Females in the BTS group were more likely to report healthy physical activity.

The results from the Chi-Square tests and the binary logistic regressions indicate that there is only a significant relationship between being in the BTS group or control group, and self-reported non-healthy physical activity, but not for the other dependent variables. This effect is negative, which is in the opposite direction from what was predicted. Respondents in the BTS group report higher amounts of healthy physical activity, in comparison to the control group. Therefore, the hypothesis that respondents in the BTS condition are more likely to report non-healthy behaviour compared to respondents in the control condition, is rejected.

Variable	Eat	Fruit	Vegetable	PA	Moderate	Intensive
BTS (1)	0.445	-0.326	-0.095	-1.067	-0.189	-0.095
	(0.261)	(0.235)	(0.731)	(0.001)***	(0.490)	(0.734)
BMI	-0.126	-0.029	-0.023	-0.069	-0.027	0.006
	(0.007)**	(0.377)	(0.482)	(0.059)	(0.408)	(0.864)
Age	0.061	0.012	-0.016	0.026	0.018	-0.020
	(0.007)**	(0.263)	(0.138)	(0.057)	(0.104)	(0.072)
Education	(0.846)	(0.470)	(0.167)	(0.942)	(0.184)	(0.126)
2	19.056	0.378	1.636	0.294	-0.209	1.601
	(0.999)	(0.654)	(0.082)	(0.770)	(0.802)	(0.086)
3	0.308	-0.001	0..175	0.307	-0.214	0.227
	(0.643)	(0.999)	(0.699)	(0.543)	(0.634)	(0.616)
4	-0.074	0.443	0.591	0.278	0.433	0.703
	(0.906)	(0.330)	(0.189)	(0.577)	(0.333)	(0.119)
Constant	2.763	-0.035	0.964	2.143	0.145	0.390
	(0.026)	(0.967)	(0.258)	(0.025)	(0.864)	(0.649)
Nagelkerke R^2	0.142	0.031	0.051	0.103	0.043	0.055
χ^2	19.028	5.261	8.630	16.701	7.334	9.397
df	6	6	6	6	6	6
Sig.	0.004**	0.511	0.195	0.010*	0.291	0.152
Classification Table Overall %	85.3	60.3	61.6	70.1	58.9	57.1

Table 7 B*inominal logistic regressions for the prediction of healthy general eating behaviour, fruit consumption, vegetable consumption, general physical activity, moderate physical activity, and intensive physical activity. In parentheses, p-values are given, *** indicates a Bonferroni significance level of 0.01/6 = 0.0017, ** indicates a Bonferroni significance level of 0.05/6= 0.008, * indicates a Bonferroni significance level of 0.10/6=0.017.*

5.3 BTS and Weight Watching

In this section the second hypothesis, the difference in responses to healthy behaviour questions between the BTS and the control group is higher for weight watchers than for non-weight watchers, is examined. The percentage of females who self-report non-healthy behaviour is graphically illustrated in Figure 3. Whether the differences between the two groups are significant was examined by six log linear analyses and six binary logistic regressions, which will be discussed in the following paragraphs.

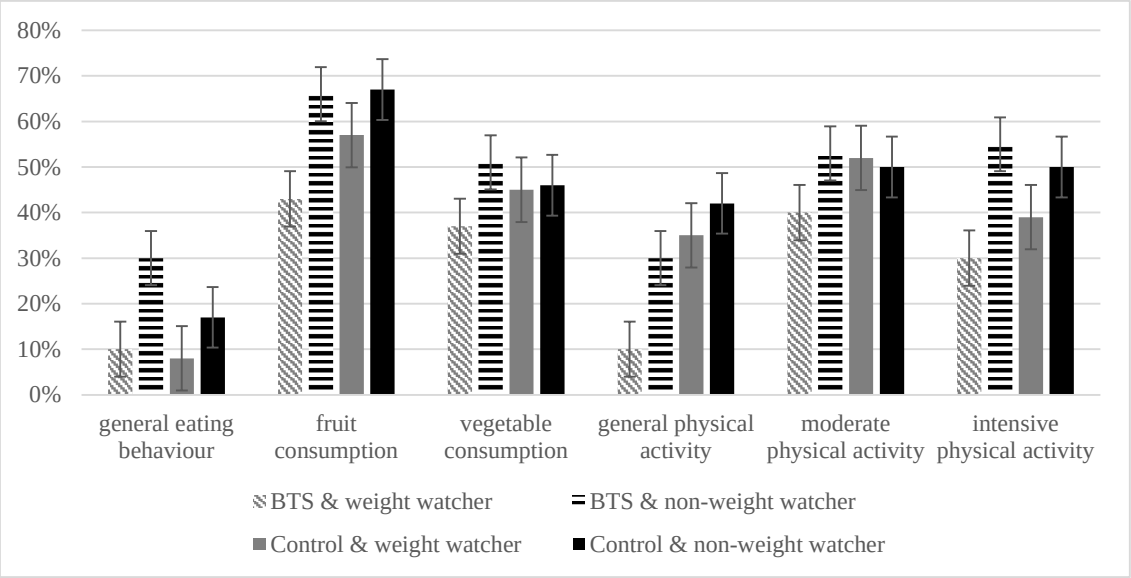

Figure 3: *Self-reporting non-healthy behaviour by group*

For both eating behaviour and physical activity three log linear analyses were conducted. These log linear analyses were used to see whether there was an interaction effect of BTS, Weight watcher, and a certain type of behaviour. Each log linear analysis includes a crosstabulation table (Appendix E, Tables 1, 3, 5, 7, 9). These crosstabulation tables were used to check the assumption for the log linear analyses. From the tables no expected counts less than five could be found. Therefore the expected frequencies are large enough to have reliable analyses. Furthermore, each respondent could only belong to one of the four categories, therefore, the assumptions of a log linear analysis hold. The results of the log linear analyses can be found in Appendix E.

In the first log linear model general eating behaviour was examined. The model includes the main effects of BTS, weight watcher, and general eating behaviour, as well as the two-way interactions of these variables; BTS*WeighWatcher, BTS*Eat, WeightWatcher*Eat. Excluding the three-way interaction (BTS*WeightWatcher*Eat) from the model does not statistically affect the model (*LR*: 0.601, χ^2 = 0.276). Therefore, the best estimated model only included the main effects and the two-way interaction effects. The main effect of 'eat' is significant, $\chi^2(1) = 119.774, p = 0.000,$ indicating (based on the contingency table) that significantly more respondents answered to generally eat healthy (190) than to not eat healthy in general (34). There is also a significant association between being a weight watcher or a non-weight watcher and whether someone self-reported healthy general eating behaviour, $\chi^2(1) = 8.938, p = 0.003, \varphi = 0.200$. Weight watchers are more likely to self-reporting healthy general eating behaviour than for non-weight watchers. The second log linear model is the one

for fruit consumption, this model only includes the main effects. The main effect of fruit consumption is significant $\chi^2(1) = 5.181, p = 0.023$ indicating (based on the contingency table) that significantly less respondents self-reported to consume 250 grams of fruit for at least five days a week (95) than to consume fewer grams (129). The third log linear model is for vegetable consumption. It appears that the best model for vegetable consumption does not include any of the effects $\chi^2(7) = 8.881, p = 0.261$. Therefore no results can be obtained for this particular question. All in all, the results of the log linear analyses for eating behaviour do not find an interaction effect of the BTS, the self-reported behaviour, and whether someone is watching her weight or not.

Also for physical activity log linear analyses were conducted. The fourth log linear analysis is the three-way log linear analysis for general physical activity. This analysis produced a final model that retained the main effects and the interaction terms BTS*PA, and WeightWatcher*PA The likelihood ratio of this model was $\chi^2(2) = 2.788, p = 0.248$. The main effect of PA is statistically significant $\chi^2(1) = 40.694, p = 0.000$ indicating (based on the contingency table) that significantly more respondents self-reported to perform general healthy physical activity (159) than to not self-report this (65). Furthermore, there is a significant association between BTS and self-reported physical activity, $\chi^2(1) = 10.605, p = 0.001, \varphi_V = 0.218$. Indicating that respondents in the BTS group have a higher likelihood to report healthy general physical activity, compared to respondents in the control group. Also, there is a significant association between whether someone is a weight watcher or not and the self-reported physical activity $\chi^2(1) = 4.648, p = 0.031, \varphi_V = 0.144$. Indicating that weight-watchers are more likely to self-report healthy physical activity, compared to non-weight watchers. The fifth log linear analysis was conducted for moderate physical activity. The best estimated log linear model for moderate physical activity, appears not to include any of the effects $\chi^2(7) = 6.224, p = 0.514$. The sixth log linear model was estimated for intensive physical activity. This model only consists of the main effects. The main effect of intensive physical activity is significant $\chi^2(1) = 5.181, p = 0.023$ indicating (based on the contingency table) that significantly more females self-reported to perform healthy intensive physical activity at least two days a week (129) than to exercise less (95). All together, no interaction between healthy physical activity, BTS, and weight watching was found.

These results suggest that there is no interaction effect of BTS, weight watcher, and any of the healthy behaviour types. However, some other results were found. For general eating behaviour it was found that weight watchers are more likely to report healthy behaviour

compared to non-weight watchers. Next, respondents in the BTS group were more likely to report healthy general physical activity, compared to the control group. Furthermore, weight watchers are more likely to report healthy general physical activity. Therefore, there is some support for the relationship between being conscious about your weight and self-reported behaviour. However, the BTS and whether someone was categorised as a weight watcher did not have a combined effect of self-reported behaviour.

The log linear analyses do not take into account any of the demographical variables, and therefore also binary logistic regressions were conducted. Six logistic regressions were performed to ascertain the effect of the four categories, BMI, age, and education on the likelihood that a respondent will self-report healthy behaviour. Since BMI and age are continuous variables, it was checked whether the variables were linearly related to their log. The results of these tests are shown in Appendix E Table 13; these indicate that this assumption is satisfied. The results of the final binary logistic regressions can be found in Table 8.

Six binary logistic regressions were conducted, three for eating behaviour and three for physical activity. The first logistic regression is the one that examines the relationship of BTS_WeightWatcher, age, BMI, and education on the likelihood to self-report healthy general eating behaviour. This model was found to be statistically significant, $\chi^2(8) = 31.996, p = 0.000$. The model explained 23.2% of the variance in self-reported general eating behaviour, and classified 87.9% of the cases correctly. This prediction is more accurate than the toss of a coin. The combined effect of the four groups of BTS_WeightWatcher is significant ($p = 0.004$), however, the separate groups are not. Furthermore, the variable BMI is significant ($p = 0.001$), indicating that respondents with a higher BMI are less likely to report healthy general eating behaviour. The variable age is also significant ($p = 0.009$), females who are older are more likely to report healthy general eating behaviour. The second binary logistic regression is the one of BTS_WeightWatcher, age, BMI, and education on the likelihood to report healthy fruit consumption. This regression was not statistically significant, $\chi^2(8) = 13.207, p = 0.105$. The model explained 7.7% of the variance in reported healthy fruit consumption, and correctly classified 62.1% of the cases. Third, the binary logistic regression of BTS_WeightWatcher, age, BMI, and education on the likelihood to self-report healthy vegetable consumption was not statistically significant, $\chi^2(8) = 11.662, p = 0.167$. The model explained 6.8% of the variance in self-reported healthy vegetable consumption, and correctly classified 58.9% of the cases. Taken together, only a combined effect of BTS and weight watcher is found for general

eating behaviour. However, the individual groups did not significantly affect the self-reported behaviour. Whether someone was in the BTS group and a non-weight watcher did significantly influence the self-reported fruit consumption. Someone in this group was more likely to report healthy fruit consumption.

For physical activity the logistic regressions were also performed. The fourth binary logistic regression is the one for general physical activity. This regression is significant, $\chi^2(8) = 22.663, p = 0.004$. The model explained 13.7% of the variance in self-reported physical activity, and correctly classified 71.9% of the cases. The combined effect of BTS_WeightWatcher is significant (p = 0.003). Females who are in the BTS group and are weight watchers are more likely to self-report healthy physical activity, compared to females who are in the control group and are non-weight watchers. The fifth binary logistic regression examines the effect for moderate physical activity, this was found to be insignificant, $\chi^2(8) = 9.684, p = 0.288$. The model explained only 5.6% of the variance is self-reported moderate physical activity, and correctly classified 58% of the cases. Sixth, the binary logistic regression of intensive physical activity is significant $\chi^2(8) = 21.120, p = 0.007$. The model explained 12.1% of the variance in self-reported intensive physical activity, and correctly classified 63.4% of the cases. The combined effect of BTS_WeightWatcher is significant (p = 0.011). Females who are in the BTS group and are weight watchers are more likely to report healthy intensive physical activity, compared to females who are non-weight watchers in the control group (p = 0.008). All in all, for general physical activity and intensive physical activity a combined effect of BTS and weight watcher was found. Weight watchers in the BTS group are more likely to report healthy physical activity and healthy intensive physical activity.

In summary, no direct interaction effect of BTS and weight watching on the different self-reported behaviours was found. However, other interesting effects were found. For general eating behaviour it was found from the log linear analysis that weight watchers are more likely to report healthy eating behaviour. This was confirmed by the results of the binary logistic regression in which the combined effect of BTS and weight watching had an effect. A similar effect was found for fruit consumption, here also females in the BTS and weight watchers were more likely to report healthy behaviour. In both analyses it was found that weight watchers in the BTS group are more likely to report healthy physical activity in general and healthy intensive physical activity. It was predicted that this group would report less healthy behaviour which was not found. Furthermore, only for weight watchers in the BTS group a significant

effect was found. Therefore, the hypothesis that the difference in responses to healthy behaviour questions between the BTS and the control group is higher for weight watcher than for non-weight watchers is rejected.

	Eat	Fruit	Vegetable	PA	Moderate	Intensive
BTS_Weight Watcher	(0.004)**	(0.029)	(0.376)	(0.003)**	(0.426)	(0.011)*
BTS* WeightWatcher	1.010 (0.099)	1.059 (0.009)*	0.480 (0.232)	1.885 (0.000)***	0.377 (0.341)	1.113 (0.008)**
BTS*Non WeightWatcher	-0.793 (0.134)	0.003 (0.994)	-0.176 (0.672)	0.516 (0.240)	-0.222 (0.591)	-0.170 (0.684)
Control* WeightWatcher	1.054 (0.090)	0.480 (0.227)	0.317 (0.416)	0.381 (0.339)	-0.148 (0.701)	0.630 (0.110)
BMI	-0.169 (0.001)***	-0.050 (0.152)	-0.035 (0.301)	-0.086 (0.022)	-0.036 (0.281)	-0.018 (0.610)
Age	0.060 (0.009)*	0.013 (0.262)	-0.017 (0.126)	0.026 (0.060)	0.019 (0.086)	-0.022 (0.054)
Education	(0.928)	(0.431)	(0.135)	(0.946)	(0.182)	(0.075)
2	19.4881 (0.999)	0.570 (0.511)	1.807 (0.062)	0.408 (0.698)	-0.228 (0.787)	2.027 (0.041)
3	0.288 (0.689)	0.024 (0.960)	0.179 (0.696)	0.294 (0.568)	-0.235 (0.605)	0.283 (0.545)
4	-0.137 (0.841)	0.477 (0.306)	0.603 (0.186)	0.268 (0.598)	0.420 (0.352)	0.769 (0.100)
Constant	3.815 (0.006)**	-0.181 (0.841)	1.028 (0.246)	1.318 (0.179)	0.232 (0.791)	0.483 (0.593)
Nagelkerke R^2	0.232	0.077	0.068	0.137	0.056	0.121
χ^2	31.996	13.207	11.662	22.663	9.684	21.120
df	8	8	8	8	8	8
Sig.	0.000	0.105	0.167	0.004	0.288	0.007
Classification Table Overall %	87.9	62.1	58.9	71.9	58.0	63.4

Table 8: B*inominal logistic regressions for the prediction of healthy general eating behaviour, fruit consumption, vegetable consumption, general physical activity, moderate physical activity, and intensive physical activity. In parentheses, p-values are given, *** indicates a Bonferroni significance level of 0.01/6 = 0.0017, ** indicates a Bonferroni significance level of 0.05/6= 0.008, * indicates a Bonferroni significance level of 0.10/6=0.017.*

6 Discussion

The central focus of this paper was to examine whether the Bayesian Truth Serum (BTS) could elicit more honest and thoughtful self-reported responses to healthy behaviour questions. Six healthy behaviour questions were used regarding eating behaviour and physical activity. Specifically, we asked respondents about their general eating behaviour, fruit consumption, vegetable consumption, general physical activity, moderate physical activity, and intensive physical activity. Previous research indicates that the BTS elicits more honest and thoughtful response to sensitive questions (John et al., 2012). Furthermore, people normally under-report eating behaviour (Hill & Davies, 2001; Schoeller, 1990; Trijsburg et al., 2016), and over-report physical activity (Rzewnicki, Auweele, and De Boudeaudhuij, 2003; Van de Mortel, 2008). Therefore it was expected that respondents in the BTS group would report less healthy behaviour (i.e., less healthy food consumption and less physical activity), compared to respondents in the control group.

The first hypothesis regards whether females that are exposed to the BTS have a higher likelihood of reporting non-healthy behaviour, compared to females that are not exposed to the BTS. This was tested by six Pearson Chi-Square tests and six binary logistic regressions. From these tests it was found that females in the BTS group are more likely to report healthy physical activity, compared to females in the control group. This implies that the opposite effect of BTS on self-reported healthy behaviour is found since we expected that fewer respondents would self-report healthy behaviour in the BTS group. For the other questions raised in the experiment, no statistical difference was found between the control and treatment group. These results suggest that overall, the BTS does not seem to elicit more honest and thoughtful self-reported responses to healthy behaviour questions in general. Interestingly, it seems that the BTS does elicit less honest self-reported responses to general physical activity questions.

The second hypothesis examines whether the BTS elicits a stronger response from weight watchers compared to non-weight watchers. We expected that such an interaction effect would exist since people who are trying to lose weight are even more likely to misreport healthy behaviour (Lichtman et al., 1992; Muhlheim, Allison, Heshka, & Heymsfield, 1998). Such an interaction effect was not directly found. However, other interesting effects were found. For general eating behaviour weight watchers were more likely to report healthy behaviour. Furthermore, weight watchers in the BTS group were more likely to report healthy physical activity. This effect is similar to the effect found for the first hypothesis and is also unexpected. This result suggests that the BTS did elicit less reliable responses about physical activity.

Furthermore, weight watchers are reporting more healthy behaviour. People who are conscious about their weight tend to report behaving more healthily, this result is in line with Lichtman et al. (1992). Consequently, this study confirms that it is crucial to increase the reliability of self-reports, especially for people who are watching their weight.

Another interesting result was that the questions about behavioural in general were easier to predict in comparison to the more specific questions. The questions about general eating behaviour and general physical activity had the highest accuracy rates. In earlier research it was already found that one single question can give a good estimation of physical activity (Schechtman et al., 1991). Also, Van Beek and colleagues (2013) used the same questions as were asked. This could mean that one single general question could better predict behaviour than separate specific questions as was done for fruit consumption and vegetable consumption.

Taken together, these findings suggest that the BTS does not elicit more honest and thoughtful self-reported responses to healthy behaviour questions in general. That is, given that the more honest and thoughtful answers would mean the non-healthy answers. The outcome of this study is surprising since current literature indicates that the BTS does elicit more honest answers (Barrage, & Lee, 2010; John et al., 2012; Loughran, Paternoster, & Thomas, 2014; Miller, Bailey, & Kirlik, 2014; Prelec, 2004). However, we are not the first to find no effect of the BTS. Menapace and Raffaelli (2015), also did not find a significant effect of the BTS. They implemented the BTS on a choice experiment about Italian pasta. These researchers also used the BTS to try to overcome the social desirability bias, which might not have worked. Although this is the only published evidence where the BTS does not seem to have an effect, this might be due to the publication bias (Rothstein, Sutton & Borenstein, 2006). Research in which no significant effect is found, is less likely to be published. It could be therefore possible that the BTS does not have an effect on healthy behaviour questions.

Another possibility why no effect was found for some questions might be because the questions were not sensitive enough and therefore the respondents already reported honestly. Loughran and colleagues (2014) examined the BTS in the field of criminology. They found an effect of the BTS on some questions, but not on other questions. The BTS elicited more reliable responses when respondents were asked about driving drunk and cheating on an exam. However, answers to questions about texting while driving and smoking marijuana did not differ between the BTS and control group. Loughran and colleagues suggest that people are more honest in general about behaviours that are less serious and more socially acceptable. Therefore, respondents in the control group would already answer honestly about texting while

driving and smoking marijuana, resulting in a small difference between the BTS group and the control group. For behaviours that are more serious and less socially acceptable people would be less honest in general. This would result in less honest answers about driving drunk and cheating on an exam in the control group. The difference between the control group and the BTS group is significant because of the effect of the BTS. For the current study this could indicate that no difference between the BTS and control group was found because the behaviour in question was less serious and more socially acceptable. Previous research on the BTS in other fields, for example unethical research practices (John et al., 2012), might be less socially acceptable compared to healthy behaviour. Therefore, it could be that the BTS did not work for healthy behaviour questions. Whether the BTS did not have an effect due to the type of questions that were asked could be examined. The effect of the BTS might have an effect for more sensitive questions about healthy behaviour. Questions that are less socially acceptable could be studied to examine this, for example questions about binge eating.

For physical activity an effect of the BTS was found, however, in the unexpected direction. Respondents reported to behave more healthy in the BTS group. Based on the over-reporting of physical activity (Rzewnicki, Auweele, and De Boudeaudhuij, 2003; Van de Mortel, 2008), we expected females in the BTS group, who were incentivised to answer more truthfully, to report less healthy physical activity compared to respondents in the control group. There are several possible explanations for this unexpected finding. First, the most straightforward reason is that the common prior assumption did not hold for the physical activity questions. As explained in section 5.1 the results of the independent t-test indicated that the common prior assumption does not hold for the question of general physical activity. Females self-reporting to perform healthy physical activity do not estimate a significantly higher percentage of other females to also perform healthy physical activity, compared to females who did not report to perform healthy physical activity. The common prior assumption is crucial for the BTS to work and since it does not hold for this question, the results of this question are not reliable. This might be the reason why the opposite effect was found. Second, the results for the physical activity question might be caused by under-reporting specifically by the control group. While the majority of studies found that people over-report physical activity (Adams et al., 2005; Rzewnicki, Auweele, & De Boudeaudhuij, 2003; Smith, Duffy, & Hilton, 1999). Prince et al. (2008) found that this is not always happening. They examined the relationship between self-reports and objective measures of physical activity by comparing different studies. Here, they did not find a clear pattern of whether respondents, in general,

would over-report or under-report their physical activity. The inconsistent results might have been due to the impact of measurement methods on the objectively estimated physical activity. They suggested that more accurate measures are needed to study physical activity to see whether respondents over-report or under-report. Another study confirms the dependence on the objective measurement of physical activity in order to estimate whether self-reports lead to under-reporting or over-reporting. Milton, Clemes, and Bull (2013) examined the use of a single question in measuring physical activity. They compared a single question to the results of the accelerometers in measuring moderate to vigorous intensity physical activity (MVPA), when all minutes were included, and when it was included in clusters of ten minutes. The researchers found that a single question did correlate with the results from the accelerometers when all minutes were included. However, when all objectively measured MVPA were used, respondents under-reported their physical activity. This means that one single question could be a valid measure of physical activity, but it could also be that respondents are under-reporting their physical activity. The objective measure of physical activity has a significant effect on whether it will be concluded whether someone is over-reporting or under-reporting. The findings of Milton et al. (2013) could indicate that respondents in the current study under-reported physical activity in the control group and that respondents in the BTS group answered more honestly, and therefore reported to perform more healthy general activity. These findings suggest that the expected over-reporting of physical activity (Rzewnicki, Auweele, and De Boudeaudhuij, 2003; Van de Mortel, 2008) might be estimated based on unreliable measures that do not reflect the real physical activity.

The results of the current study also need to be interpreted with caution due to several limitations of the study. First, it is possible that no effect of the BTS was found because subjects in this study were relatives and friends from the experimenter. This might have had several unintended effects. For example, social norms influence eating behaviour as was found in several studies (Croker et al., 2009; Eric et al., 2014; Higgs & Thomas, 2016; Robinson et al., 2011) Future research could use a randomly selected sample of people who are not related to the researcher. Or it could be tested whether knowing the experimenter will influence the results by creating a sample consisting of strangers and friends. Second, the given incentive during the experiment is only 50 euros for the whole sample. John et al. (2012) had a budget of 2,000 euro. This is a large difference, and the small incentive in the current study might not have incentivised respondent. Third, the rather small sample size might be the reason for not finding an effect of the BTS. The sample only consists of 224 respondents. This is a small sample since

John et al. (2012) used a sample of 2,155 respondents. However, the sample size of Loughran, et al. (2014) was smaller, it consisted of 137 respondents. They did find a significant effect of the BTS on some questions.

A fourth possible limitation of the study is the way in which the BTS-follow-up questions were phrased. While we tried to minimise the social desirability bias by using the BTS, the way these questions were phrased might have accidentally increased the bias. For example, it was stated that it would be recommended by the Dutch government to exercise intensively for 30 minutes at least two times a week. How this is phrased could have induced the social desirability bias instead of decreasing it. By stating the recommendations of the government also social norms are made clear, people are mostly not aware of social norms but it was found that these influence our behaviour (Croker et al., 2009). These effect of social norms might have increased the social desirability bias. The probability that social desirability might have been increased instead of decreased is of concern since the BTS was used to try to overcome this bias. In order to test this in a follow-up study, social desirability bias could be measured separately by using the Crowne and Marlowe (1960) social desirability scale. It could then be examined how the respondent's score on this scale relates to the answers of healthy behaviour questions.

The fifth possible limitation could be found in the design of the questionnaire. The BTS-questionnaire consisted of 26 questions, and the Control-questionnaire consisted of only 18 questions. This could have let to more questionnaire fatigue in the BTS group since questionnaire fatigue increases when the length of the questionnaire increases (Galesic & Bosnjak, 2009). Since the drop-out rate in the BTS group is also higher (7.76%) than the drop-out rate in the Control group (0%), this is likely to be the case. One way to solve this problem is the use of filler questions (Malhotra, 2006). The validity of the questionnaire could be improved by adding filler questions to the control group's questionnaire. By doing this, the BTS and control questionnaire become of equal length, and respondents are equally likely to suffer from questionnaire fatigue. The filler questions should be similar to the BTS-questions, however, not related to the healthy behavioural question. For example, subjects could be asked to estimate the percentage of other subjects in the experiment that is above the age of 20.

Overall, this study suggests that there is no significant general effect of the BTS on self-reported healthy behaviour for eating behaviour, fruit consumption, vegetable consumption, moderate physical activity, and intensive physical activity. However, an effect of BTS on general physical activity was found. More precisely, the BTS made females report more healthy

general physical activity compared to the control group. It must be borne in mind that this study used a small sample that was related to the researcher and that respondents in the BTS group might have suffered from questionnaire fatigue and the social desirability bias. Furthermore, self-reported healthy behaviour needs to be investigated because this could be an efficient way to measure whether incentives to live a healthier lifestyle do work. Therefore, more research is necessary to find out whether the BTS can elicit more honest and thoughtful responses in healthy behaviour questions. The abovementioned suggestions can be implemented in the new study design to enhance the current design and increase its reliability. Further research could also examine alternative ways to implement the BTS on questionnaires since other research suggests it might work for other sensitive topics.

Appendix A: Questionnaire

A.1 Introduction Text

Beste geïnteresseerde,

Hartelijk bedankt voor uw interesse en deelname in dit onderzoek. Allereerst wil ik u melden dat dit onderzoek alleen gericht is op vrouwen. Dit onderzoek doe ik voor mijn Masterscriptie voor de opleiding Business Economics, Behavioural Economics aan de Erasmus Universiteit Rotterdam. De afname van de enquête duurt ongeveer drie tot zeven minuten. Uw antwoorden op de vragen zijn **anoniem** en kunnen dus op geen enkele manier tot u worden herleid. Omdat ik u vanwege uw gegarandeerde anonimeit niet zelf kan belonen, zal ik namens u en alle andere respondenten €50 doneren verspreid over verschillende goede doelen.

Wanneer u nog vragen of opmerkingen heeft over het onderzoek kunt u contact met mij opnemen via onderstaand mail adres: 413155kw@student.eur.nl

Nogmaals hartelijk bedankt voor uw deelname aan de vragenlijst!

Met vriendelijke groet,

Karlijn de Wilde

A.2 Sample Control Questions

De volgende vragen zijn er om te bepalen of u tot de doelgroep behoort.

Wat is uw geslacht?

- ○ vrouw
- ○ man

Lijdt u op dit moment aan de eetstoornis Anorexia Nervosa?

- ○ nee
- ○ ja

Lijdt u op dit moment aan de eetstoornis Boulimia Nervosa?

- ○ nee
- ○ ja

A.3 General Questions

Goed nieuws, u behoort tot de doelgroep! Nu volgen er een aantal algemene vragen, ik wil daarbij nogmaals benadrukken dat u gegevens anoniem zijn.

Wat is uw leeftijd in jaren?

_____

Wat is uw hoogst behaalde onderwijsniveau?

- ○ Basisonderwijs
- ○ Vmbo, mbo 1, avo, onderbouw
- ○ Havo, vwo, mbo
- ○ Hbo-bachelor, wo-bachelor
- ○ Hbo-master, wo-master, doctor

Wat is uw lengte in centimeters?

_____

Wat is uw gewicht in kilogrammen?

_____

Bent u op dit moment zwanger?

- ○ nee
- ○ ja

Geef aan met welk van de volgende stellingen u zich momenteel het best kunt identificeren:

- ○ Ik probeer actief gewicht te verliezen
- ○ Ik probeer actief op gewicht te blijven
- ○ Ik probeer actief aan te komen
- ○ Ik ben niet actief met mijn gewicht bezig

Hoe tevreden bent u met uw huidige lichaamsgewicht?

- erg tevreden
- tevreden
- niet tevreden, niet ontevreden
- ontevreden
- erg ontevreden

Heeft u **de afgelopen twee weken** een of meerdere van de volgende maatregelen genomen om gewicht te verliezen of om ervoor te zorgen dat u niet verder aan kwam?

	ja	nee
Meer - matig intensief bewogen (zoals wandelen)	)	)
Meer - intensief bewogen (zoals hardlopen)	)	)
Meer fruit gegeten	)	)
Meer groente gegeten	)	)

Hoe goed is het de afgelopen twee weken gelukt om de voorgenomen maatregelen vol te houden?

- zeer voldoende
- voldoende
- niet voldoende, niet onvoldoende
- onvoldoende
- zeer onvoldoende
- niet van toepassing

A.4 Control Instruction

Bedankt voor het invullen van deze vragen, er zullen nog zes vragen volgen. Om u te bedanken voor uw deelname zal ik geld doneren aan de volgende goede doelen:

UNICEF (https://www.unicef.nl)

Artsen Zonder Grenzen (https://artsenzondergrenzen.nl/)

The Donkey Sancturary (https://www.donkeysanctuary.nl/)

Het totale bedrag dat aan deze goede doelen te samen wordt gedoneerd is €50. De donatie zal gedaan worden namens alle respondenten.

A.5 BTS Instruction

Bedankt voor het invullen van deze vragen, er zullen nog veertien vragen volgen. Om u te bedanken voor uw deelname zal ik €50 doneren verspreid over de volgende goede doelen:

UNICEF (https://www.unicef.nl)
Artsen Zonder Grenzen (https://artsenzondergrenzen.nl/)
The Donkey Sancturary (https://www.donkeysanctuary.nl/)

U kunt invloed uitoefenen op de verdeling van het geld over de goede doelen, dit doet u door zo eerlijk mogelijk antwoord te geven!

Een formule, uitgevonden door een MIT professor, zal gebruikt worden om objectief te bepalen hoe eerlijk u geantwoord heeft op de vragen. Deze formule heet het Bayesian Truth Serum en het beloond het geven van eerlijke antwoorden. Wanneer u alle vragen eerlijk beantwoord, heeft uw voorkeur voor een bepaald goed doel meer invloed op de uiteindelijke verdeling. Het is niet belangrijk voor u om te weten hoe deze formule werkt maar mocht u geïnteresseerd zijn dan kunt u het wetenschappelijke artikel dat gepubliceerd is in *Science* hier lezen.

Dit kan er gedaan worden met het geld wanneer het totale bedrag gaat naar het goede doel van uw voorkeur:

UNICEF:
"Voor €50,- kan UNICEF vaccins aanschaffen om 147 kinderen te beschermen tegen polio" (https://www.unicef.nl/doneren)

Artsen Zonder Grenzen:
"Met €50 kunnen wij ACT-combinatiepillenkuren kopen om 80 volwassenen binnen 3 dagen van malaria te genezen." (https://form.artsenzondergrenzen.nl/doe-een-gift)

The Donkey Sanctuary:
"Met €50,- kunnen wij de voorraad medicijnen voor 1 dag bijvullen, om zieke en gewonde ezels wereldwijd te behandelen" (https://www.donkeysanctuary.nl/manieren-om-te-helpen)

Om te controleren of u de intentie van het Bayesian Truth Serum begrijpt vraag ik u de volgende vraag te beantwoorden.

Het geven van eerlijke antwoorden op deze vragenlijst….. mijn invloed op de verdeling van het geld over de goede doelen.

- ○ verandert niets aan
- ○ vergroot
- ○ verlaagt

Aan welk goed doel wilt u dat het geld gedoneerd wordt?

- ○ UNICEF
- ○ Artsen zonder Grenzen
- ○ The Donkey Sanctuary

A.6 Healthy Behaviour Questions

De volgende vraag heeft betrekking op uw eetgedrag van de afgelopen 2 weken. Ik doe dus een beroep op uw geheugen. [1]

1) Bent u het eens met de volgende stelling: “Over het algemeen eet ik gezond.”

- ○ ja
- ○ nee

2) Geef aan op hoeveel dagen van de week u ten minste 200 gram (ongeveer twee porties) fruit heeft gegeten:

0 1 2 3 4 5 6 7

Geef aan op hoeveel dagen:

[1] This instruction was shown before every question in this section

3) Geef aan op hoeveel dagen van de week u ten minste 250 gram groente heeft gegeten:

0 1 2 3 4 5 6 7

Geef aan op hoeveel dagen: 

4) Bent u het eens met de volgende stelling: “Over het algemeen beweeg ik genoeg.”

○ ja

○ nee

5) Geef aan op hoeveel dagen van de week u ten minste 30 minuten matig intensief heeft bewogen (bijvoorbeeld lopen):

0 1 2 3 4 5 6 7

Geef aan op hoeveel dagen: 

6) Geef aan op hoeveel dagen van de week u ten minste 20 minuten intensief heeft bewogen

0 1 2 3 4 5 6 7

Geef aan op hoeveel dagen: 

A.7 BTS Questions

1) Hoeveel procent van de andere respondenten in het onderzoek eet volgens u over het algemeen gezond?

Geef aan hoeveel procent: (1-99%) 

2) Het wordt aanbevolen om minimaal 200 gram (ongeveer twee porties) fruit per dag te eten. Hoeveel procent van de andere respondenten denkt u dat dit 7 dagen per week doet?

Geef aan hoeveel procent: (1-99%) 

3) Het wordt aanbevolen om minimaal 250 gram groente per dag te eten. Hoeveel procent van de andere respondenten denkt u dat dit 7 dagen per week doet?

Geef aan hoeveel procent: (1-99%) 

4) Hoeveel procent van de andere respondenten in het onderzoek denkt u dat over het algemeen genoeg beweegt?

Geef aan hoeveel procent: (1-99%) 

5) Hoeveel procent van de andere respondenten denkt u dat ten minste 5 dagen in de week 30 minuten matig intensief beweegt per week?

Geef aan hoeveel procent: (1-99%) 

6) Hoeveel procent van de andere respondenten denkt u dat ten minste 3 dagen in de week 20 minuten intensief beweegt per week?

Geef aan hoeveel procent: (1-99%) 

Appendix B: Results tests common prior assumption

Variable	Obs	Prob>z
Prediction_eat	107	0.027**
Prediction_fruit	107	0.317
Prediction_vegetable	107	0.188
Prediction_PA	107	0.607
Prediction_moderate	107	0.989
Prediction_intensive	107	0.512

Table 1: *Shapiro-Wilk test*

Variable	sdtest			Robvar		
				W0	W50	W10
	Pr(F<f)	2*Pr(F<f)	Pr(F>f)	Pr>F	Pr>F	Pr>F
Prediction_eat	0.363	0.727	0.637	0.901	0.937	0.979
Prediction_fruit	0.203	0.405	0.797	0.802	0.793	0.791
Prediction_vegetable	0.893	0.214	0.107	0.154	0.151	0.149
Prediction_PA	0.229	0.458	0.771	0.224	0.260	0.233
Prediction_moderate	0.155	0.310	0.845	0.485	0.453	0.476
Prediction_intensive	0.274	0.548	0.726	0.494	0.486	0.490

Table 2: *Levene's test*

eat	Obs	Rank sum	Expected
Non-healthy eating behaviour	20	781	1080
Healthy eating behaviour	87	4997	4698
Z	-2.393		
Prob > IzI	0.0167		
$\eta^2 = \frac{Z^2}{N-1}$	0.054		

Table 3: *Two-sample Wilcoxon rank-sum (Mann-Whitney) test for the prediction of eating behaviour*

Group	Obs	Mean	Std. Err.	Std. Dev.
Non-healthy	57	39.772	2.314	17.468
Healthy	50	49.720	2.770	19.590
Combined	107	44.421	1.843	19.065
Diff		-9.948	3.582	
Ha: diff < 0	0.003			
Ha: diff != 0	0.007			
Ha: diff > 0	0.997			
$d = \frac{M_1 - M_2}{SD_{pooled}}$	0.536			

Table 4: *t-test for the estimation of fruit consumption by consuming a healthy amount of fruit*

Group	Obs	Mean	Std. Err	Std. Dev
Non-healthy	46	42.478	3.073	20.845
Healthy	61	55.934	2.249	17.567
Combined	107	50.150	1.943	20.098
Diff		-13.46	3.718	
Ha: diff < 0	0.0002			
Ha: diff ! = 0	0.0005			
Ha: diff > 0	0.9998			
$d = \frac{M_1 - M_2}{SD_{pooled}}$	0.698			

Table 5: *t-test for the estimation of vegetable consumption by consuming a healthy amount of vegetables*

Group	Obs	Mean	Std. Err	Std. Dev
Non-healthy	20	50.5	3.340	14.937
Healthy	87	53.276	1.865	17.399
Combined	107	52.276	1.637	16.934
Diff		52.757	4.211	
Ha: diff < 0	0.256	-2.776		
Ha: diff ! = 0	0.511			
Ha: diff > 0	0.744			
$\boldsymbol{d = \frac{M_1 - M_2}{SD_{pooled}}}$	0.171			

Table 6: *t-test for the estimation of PA by own report of PA*

Group	Obs	Mean	Std. Err	Std. Dev
Non-healthy	49	46.408	2.428	16.997
Healthy	58	63.052	2.575	19.612
Combined	107	55.430	1.951	20.176
Diff		-16.644	3.582	
Ha: diff < 0	0.000			
Ha: diff ! = 0	0.000			
Ha: diff > 0	1.000			
$\boldsymbol{d = \frac{M_1 - M_2}{SD_{pooled}}}$	0.907			

Table 7: *t-test for the estimation of moderate physical activity by own report of moderate physical activity*

Group	Obs	Mean	Std. Err	Std. Dev
Non-healthy	44	33.273	2.411	15.993
Healthy	63	44.730	2.199	17.458
Combined	107	40.019	1.713	17.723
Diff		-11.457	3.315	
Ha: diff < 0	0.0004			
Ha: diff ! = 0	0.0008			
Ha: diff > 0	0.9996			
$\boldsymbol{d = \frac{M_1 - M_2}{SD_{pooled}}}$	0.684			

Table 8: *t-test for the estimation of intensive physical activity by own report of intensive physical activity*

Appendix C: Results tests hypothesis 1

			Eating behaviour		
			Non-healthy	Healthy	Total
BTS	**Control**	**Count**	14	103	117
		% within BTS	12.0%	88.0%	100.0%
		% within eat	41.2%	54.2%	52.2%
		% of total	6.3%	46.0%	52.2%
	BTS	**Count**	20	87	107
		% within BTS	18.7%	81.3%	100.0%
		% within eat	58.8%	45.8%	47.8%
		% of total	8.9%	38.8%	47.8%

	Value	df	Sig.
Pearson Chi-Square	1.964	1	0.161
Cramer's V	0.094		

Table 1: *Pearson Chi-Square test for eating behaviour*

			Fruit Consumption		
			Non-healthy	Healthy	Total
BTS	**Control**	**Count**	72	45	117
		% within BTS	615%	38.5%	100.0%
		% within fruit	55.8%	47.4%	52.2%
		% of total	32.1%	20.1%	52.2%
	BTS	**Count**	57	50	107
		% within BTS	53.3%	46.7%	100.0%
		% within fruit	44.2%	52.6%	47.8%
		% of total	25.4%	22.3%	47.8%

	Value	df	Sig.
Pearson Chi-Square	1.564	1	0.211
Cramer's V	0.084		

Table 2: *Pearson Chi-Square test for fruit consumption*

			Vegetable Consumption		
			Non-healthy	Healthy	Total
BTS	**Control**	**Count**	53	64	117
		% within BTS	45.3%	54.7%	100.0%
		% within vegetable	53.5%	51.2%	52.2%
		% of total	23.7%	28.6%	52.2%
	BTS	**Count**	46	61	107
		% within BTS	43.0%	57.0%	100.0%
		% within vegetable	46.5%	48.8%	47.8%
		% of total	20.5%	27.2%	47.8%

	Value	df	Sig.
Pearson Chi-Square	0.121	1	0.728
Cramer's V	0.023		

Table 3: *Pearson Chi-Square test for vegetable consumption*

			General physical activity		
			Non-healthy	Healthy	Total
BTS	**Control**	**Count**	45	72	117
		% within BTS	38.5%	61.5%	100.0%
		% within PA	69.2%	45.3%	52.2%
		% of total	20.1%	32.1%	52.2%
	BTS	**Count**	20	87	107
		% within BTS	18.7%	81.3%	100.0%
		% within PA	30.8%	54.7%	47.8%
		% of total	8.9%	38.8%	47.8%

	Value	df	Sig.
Pearson Chi-Square	10.605	1	0.001
Cramer's V	0.218		

Table 4: *Pearson Chi-Square test for general physical activity*

			Moderate physical activity		
			Non-healthy	Healthy	Total
BTS	**Control**	**Count**	60	57	117
		% within BTS	51.3%	48.7%	100.0%
		% within moderate	55.0%	49.6%	52.2%
		% of total	26.8%	25.4%	52.2%
	BTS	**Count**	49	58	107
		% within BTS	45.8%	54.2%	100.0%
		% within moderate	45.0%	50.4%	47.8%
		% of total	21.9%	25.9%	47.8%

	Value	df	Sig.
Pearson Chi-Square	0.674	1	0.412
Cramer's V	0.055		

Table 5: *Pearson Chi-Square test for moderate physical activity*

			Intensive physical activity		
			Non-healthy	Healthy	Total
BTS	**Control**	**Count**	51	66	117
		% within BTS	43.6%	56.4%	100.0%
		% within intensive	53.7%	51.2%	52.2%
		% of total	22.8%	29.5%	52.2%
	BTS	**Count**	44	63	107
		% within BTS	41.1%	58.9%	100.0%
		% within intensive	46.3%	48.8%	47.8%
		% of total	19.6%	28.1%	47.8%

	Value	df	Sig.
Pearson Chi-Square	0.139	1	0.709
Cramer's V	0.025		

Table 6: *Pearson Chi-Square test for intensive physical activity*

	B	S.E.	Wald	Df	Sig.	Exp(B)
BTS	0.473	0.391	1.465	1	0.226	1.605
Education 1			1.584	3	0.663	
Education 2	18.880	13735.278	0.000	1	0.999	158376716.918
Education 3	0.346	0.652	0.281	1	0.596	1.414
LN_age	-0.193	0.612	0.099	1	0.752	0.824
LN_BMI	0.009	0.004	3.899	1	0.048	1.009
Constant	0.000	0.000	0.065	1	0.798	1.000

Table 7: *Box-Tidwell test for general eating behaviour*

	B	S.E.	Wald	Df	Sig.	Exp(B)
BTS	-0.337	0.275	1.496	1	0.221	0.714
Education 1			2.098	3	0.552	
Education 2	0.357	0.836	0.183	1	0.669	1.430
Education 3	0.028	0.462	0.004	1	0.951	1.029
LN_age	0.424	0.454	0.870	1	0.351	1.528
LN_BMI	0.002	0.002	0.650	1	0.420	1.002
Constant	0.000	0.000	0.204	1	0.652	1/000

Table 8: *Box-Tidwell test for fruit consumption*

	B	S.E.	Wald	Df	Sig.	Exp(B)
BTS	-0.116	0.277	0.176	1	0.675	0.890
Education 1			4.658	3	0.199	
Education 2	1.627	0.935	3.024	1	0.082	5.087
Education 3	0.216	0.453	0.228	1	0.633	1.241
LN_age	0.579	0.449	1.663	1	0.197	1.784
LN_BMI	-0.004	0.002	3.612	1	0.057	0.996
Constant	0.000	0.000	0.121	1	0.728	1.000

Table 9: *Box-Tidwell test for vegetable consumption*

	B	S.E.	Wald	Df	Sig.	Exp(B)
BTS	-1.081	0.321	11.343	1	0.001	0.339
Education 1			0.530	3	0.912	
Education 2	0.120	0.967	0.015	1	0.901	1.128
Education 3	0.351	0.505	0.482	1	0.487	1.420
LN_age	0.215	0.498	0.187	1	0.666	1.240
LN_BMI	0.003	0.003	1.553	1	0.213	1.003
Constant	0.000	0.000	0.301	1	0.583	1.000

Table 10: *Box-Tidwell test for physical activity*

	B	S.E.	Wald	Df	Sig.	Exp(B)
BTS	-0.200	0.274	0.535	1	0.464	0.818
Education 1			4.267	3	0.234	
Education 2	-0.233	0.832	0.078	1	0.780	0.792
Education 3	-0.188	0.450	0.175	1	0.676	0.828
LN_age	0.413	0.446	0.856	1	0.355	1.511
LN_BMI	0.003	0.002	1.921	1	0.166	1.003
Constant	0.000	0.000	0.216	1	0.642	1.000

Table 11: *Box-Tidwell test for moderate physical activity*

	B	S.E.	Wald	Df	Sig.	Exp(B)
BTS	-0.124	0.280	0.196	1	0.658	0.883
Education 1			5.599	3	0.133	
Education 2	1.636	0.936	3.054	1	0.081	5.132
Education 3	0.263	0.453	0.337	1	0.561	1.301
LN_age	0.711	0.451	2.486	1	0.115	2.035
LN_BMI	-0.004	0.002	3.751	1	0.053	0.996
Constant	0.000	0.000	0.123	1	0.726	1.000

Table 12: *Box-Tidwell test for intensive physical activity*

Appendix D: Results tests hypothesis 2

Count					
			eat		
BTS			0	1	Total
0	Weight Watcher	0	9	43	52
		1	5	60	65
	Total		14	103	117
1	Weight Watcher	0	14	33	47
		1	6	54	60
	Total		20	87	107
Total	Weight Watcher	0	23	76	99
		1	11	114	125
	Total		34	190	224

Table 1: *Log linear Analysis for general eating behaviour: WeightWatcher*Eat*BTS Crosstabulation*

			Likelihood Ratio		Pearson		Number of
	K	**df**	**Chi-Square**	**Sig.**	**Chi-Square**	**Sig**	**Iterations**
K-way and Higher Order Effects	1	7	134.576	0.000	125.714	0.000	0
	2	4	11.330	0.023	12.006	0.017	2
	3	1	0.274	0.601	0.276	0.599	3
K-way Effects	1	3	123.245	0.000	113.708	0.000	0
	2	3	11.056	0.011	11.730	0.008	0
	3	1	0.274	0.601	0.276	0.599	0

Table 2: *Log linear Analysis for general eating behaviour: K-Way and Higher-Order Effects*

Count					
			fruit5		
BTS			0	1	Total
0	Weight Watcher	0	35	17	52
		1	37	28	65
	Total		72	45	117
1	Weight Watcher	0	31	16	47
		1	26	34	60
	Total		57	50	107
Total	Weight Watcher	0	66	33	99
		1	63	62	125
	Total		129	95	224

Table 3: *Log linear Analysis for fruit consumption: WeightWatcher*Fruit5*BTS Crosstabulation*

		Likelihood Ratio		**Pearson**		**Number of**
	K	**df**	**Chi-Square**	**Sig.**	**Chi-Square**	**Sig**	**Iterations**
K-way and Higher Order Effects	1	7	17.031	0.017	15.857	0.026	0
	2	4	8.379	0.079	8.396	0.078	2
	3	1	0.757	0.384	0.757	0.384	3
K-way Effects	1	3	8.652	0.034	7.461	0.059	0
	2	3	7.622	0.055	7.639	0.054	0
	3	1	0.757	0.384	0.757	0.384	0

Table 4: *Log linear Analysis for fruit consumption: K-Way and Higher-Order Effects*

Count					
			vegetable5		
BTS			0	1	Total
0	Weight Watcher	0	24	28	52
		1	28	37	65
	Total		52	65	117
1	Weight Watcher	0	24	23	47
		1	23	37	60
	Total		47	60	107
Total	Weight Watcher	0	48	51	99
		1	51	74	125
	Total		99	125	224

Table 5: *Log linear Analysis for vegetable consumption: WeightWatcher*Vegetable5*BTS Crosstabulation*

			Likelihood Ratio		Pearson		Number of
	K	**df**	**Chi-Square**	**Sig.**	**Chi-Square**	**Sig**	**Iterations**
K-way and Higher Order Effects	1	7	8.881	0.261	9.214	0.238	0
	2	4	2.385	0.665	2.371	0.668	2
	3	1	0.939	0.333	0.938	0.333	2
K-way Effects	1	3	6.496	0.090	6.843	0.077	0
	2	3	1.446	0.695	1.433	0.698	0
	3	1	0.939	0.333	0.938	0.333	0

Table 6: *Log linear Analysis for vegetable consumption: K-Way and Higher-Order Effects*

Count					
			PA		
BTS			0	1	Total
0	Weight	0	22	30	52
	Watcher	1	22	43	65
	Total		44	73	117
1	Weight	0	14	33	47
	Watcher	1	7	53	60
	Total		21	86	107
Total	Weight	0	36	63	99
	Watcher	1	29	96	125
	Total		65	159	224

Table 7: *Log linear Analysis for general physical activity: WeightWatcher*PA*BTS Crosstabulation*

			Likelihood Ratio		Pearson		Number of
	K	**df**	**Chi-Square**	**Sig.**	**Chi-Square**	**Sig**	**Iterations**
K-way and	1	7	62.422	0.000	58.643	0.000	0
Higher Order	2	4	18.256	0.001	16.364	0.003	2
Effects	3	1	2.624	0.105	2.577	0.108	3
K-way Effects	1	3	44.166	0.000	42.279	0.000	0
	2	3	15.632	0.001	13.787	0.003	0
	3	1	2.624	0.105	2.577	0.108	0

Table 8: *Log linear Analysis for vegetable consumption: K-Way and Higher-Order Effects*

Count					
			moderate5		
BTS			0	1	Total
0	Weight	0	26	26	52
	Watcher	1	34	31	65
	Total		60	57	117
1	Weight	0	25	22	47
	Watcher	1	24	36	60
	Total		49	58	107
Total	Weight	0	51	48	99
	Watcher	1	58	67	125
	Total		109	115	224

Table 9: *Log linear Analysis for moderate physical activity: WeightWatcher*Moderate5*BTS Crosstabulation*

			Likelihood Ratio		Pearson		Number of
	K	**df**	**Chi-Square**	**Sig.**	**Chi-Square**	**Sig**	**Iterations**
K-way and	1	7	6.224	0.514	6.357	0.499	0
Higher Order	2	4	2.592	0.628	2.582	0.630	2
Effects	3	1	1.338	0.247	1.337	0.248	2
K-way Effects	1	3	3.632	0.304	3.775	0.287	0
	2	3	1.254	0.740	1.246	0.742	0
	3	1	1.338	0.247	1.337	0.248	0

Table 10: *Log linear Analysis for moderate physical activity: K-Way and Higher-Order Effects*

Count					
			intensive2		
BTS			0	1	Total
0	Weight	0	26	26	52
	Watcher	1	26	39	65
	Total		52	65	117
1	Weight	0	26	21	47
	Watcher	1	17	43	60
	Total		43	64	107
Total	Weight	0	52	47	99
	Watcher	1	43	82	125
	Total		95	129	224

Table 11: *Log linear Analysis for intensive physical activity: WeightWatcher*Intensive2*BTS Crosstabulation*

			Likelihood Ratio		Pearson		Number of
	K	**df**	**Chi-Square**	**Sig.**	**Chi-Square**	**Sig**	**Iterations**
K-way and	1	7	17.377	0.015	18.214	0.011	0
Higher Order	2	4	8.725	0.068	8.641	0.071	2
Effects	3	1	1.140	0.286	1.138	0.286	3
K-way Effects	1	3	8.652	0.034	9.573	0.023	0
	2	3	7.585	0.055	7.503	0.057	0
	3	1	1.140	0.286	1.138	0.286	0

Table 12: *Log linear Analysis for intensive physical activity: K-Way and Higher-Order Effects*

Variable	Eat	Fruit	Vegetable	PA	Moderate	Intensive
BTS_WeightWatcher	(0.007)**	(0.031)	(0.466)	(0.002)**	(0.442)	(0.015)*
BTS*NonWeightWatcher	1.076	1.073	0.438	1.922	0.352	1.096
	(0.083)	(0.009)*	(0.286)	(0.000)***	(0.376)	(0.009)*
Control*WeightWatcher	-0.676	0.053	-0.175	0.623	-0.224	-0.150
	(0.212)	(0.904)	(0.679)	(0.165)	(0.593)	(0.723)
Control*NonWeight	1.091	0.485	0.269	0.393	-0.176	0.600
Watcher	(0.082)	(0.222)	(0.497)	(0.326)	(0.647)	(0.130)
BMI	-0.186	-0.056	0.233	-0.097	-0.036	-0.026
	(0.001)***	(0.117)	(0.849)	(0.013)*	(0.424)	(0.462)
BMI*log(BMI)	0.000	0.000	-0.064	0.000	-0.001	0.000
	(0.785)	(0.667)	(0.823)	(0.653)	(0.927)	(0.529)
Age	0.105	0.236	0.018	0.228	-0.065	0.223
	(0.891)	(0.583)	(0.967)	(0.642)	(0.878)	(0.617)
Age*log(Age)	-0.009	-0.048	-0.008	-0.044	0.018	-0.053
	(0.955)	(0.603)	(0.934)	(0.681)	(0.843)	(0.581)
Education	(0.726)	(0.457)	(0.168)	(0.887)	(0.227)	(0.082)
2	19.745	0.671	1.915	0.559	-0.250	2.174
	(0.999)	(0.451)	(0.066)	(0.611)	(0.772)	(0.035)
3	0.436	0.111	0.233	0.414	-0.223	0.405
	(0.574)	(0.821)	(0.627)	(0.443)	(0.637)	(0.409)
4	-0.102	0.521	0.608	0.313	0.402	0.818
	(0.885)	(0.271)	(0.191)	(0.545)	(0380)	(0.088)
Constant	3.788	-1.726	-0.673	0.018	0.942	-1.130
	(0.495)	(0.589)	(0.510)	(0.996)	(0.765)	(0.733)
Nagelkerke R^2	0.262	0.083	0.087	0.157	0.069	0.137
χ^2	36.385	14.319	14.984	26.054	11.925	24.076
df	10	10	10	10	10	10
Sig.	0.000	0.159	0.133	0.004	0.290	0.007
Classification Table Overall %	87.9	62.1	61.2	73.2	58.9	65.2

Table 13: *Logistic Regressions to check whether the assumption of linearity holds for the variables age and BMI. In parantheses, p-values are given, *** indicates a Bonferroni significance level of 0.01/6 = 0.0017, ** indicates a Bonferroni signficicance level of 0.05/6= 0.008, * indicates a Bonferroni significance level of 0.10/6=0.017.*

References

Adams, S. A., Matthews, C. E., Ebbeling, C. B., Moore, C. G., Cunningham, J. E., Fulton, J., & Hebert, J. R. (2005). The effect of social desirability and social approval on self-reports of physical activity. *American Journal of Epidemiology*, *161*(4), 389-398.

Ainsworth, B., Cahalin, L., Buman, M., & Ross, R. (2015). The current state of physical activity assessment tools. *Progress in Cardiovascular Diseases*, *57*(4), 387-395.

Alharbi, M., Bauman, A., Neubeck, L., & Gallagher, R. (2016). Validation of Fitbit-Flex as a measure of free-living physical activity in a community-based phase III cardiac rehabilitation population. *European Journal of Preventive Cardiology*, *23*(14), 1476-1485.

Asbeck, I., Mast, M., Bierwag, A., Westenhöfer, J., Acheson, K. J., & Müller, M. J. (2002). Severe underreporting of energy intake in normal weight subjects: use of an appropriate standard and relation to restrained eating. *Public Health Nutrition*, *5*(5), 683-690.

Bandini, L. G., Schoeller, D. A., Cyr, H. N., & Dietz, W. H. (1990). Validity of reported energy intake in obese and nonobese adolescents. *The American Journal of Clinical Nutrition*, *52*(3), 421-425.

Bandura, A. (1998). Health promotion from the perspective of social cognitive theory. *Psychology and Health*, *13*(4), 623-649.

Barrage, L., & Lee, M. S. (2010). A penny for your thoughts: Inducing truth-telling in stated preference elicitation. *Economics Letters*, *106*(2), 140-142.

Black, A. E., Goldberg, G. R., Jebb, S. A., Livingstone, M. B., Cole, T. J., & Prentice, A. M. (1991). Critical evaluation of energy intake data using fundamental principles of energy physiology: 2. Evaluating the results of published surveys. *European Journal of Clinical Nutrition*, *45*(12), 583-599.

Black, A. E., Prentice, A. M., Goldberg, G. R., Jebb, S. A., Bingham, S. A., Livingstone, M. B. E., & Coward, A. (1993). Measurements of total energy expenditure provide insights into the validity of dietary measurements of energy intake. *Journal of the American Dietetic Association*, *93*(5), 572-579.

Bunt, S. N. W., Mérelle, S. Y. M., Steenhuis, I. H. M., & Kroeze, W. (2017). Predictors of need for help with weight loss among overweight and obese men and women in the Netherlands: a cross-sectional study. *BMC Health Services Research*, *17*(1), 819.

Burke, L. E., Wang, J., & Sevick, M. A. (2011). Self-monitoring in weight loss: a systematic review of the literature. *Journal of the American Dietetic Association*, *111*(1), 92-102.

Case, M. A., Burwick, H. A., Volpp, K. G., & Patel, M. S. (2015). Accuracy of smartphone applications and wearable devices for tracking physical activity data. *Jama*, *313*(6), 625-626.

Chaiken, S., & Pliner, P. (1987). Women, but not men, are what they eat: The effect of meal size and gender on perceived femininity and masculinity. *Personality and Social Psychology Bulletin*, *13*(2), 166-176.

Croker, H., Whitaker, K. L., Cooke, L., & Wardle, J. (2009). Do social norms affect intended food choice?. *Preventive Medicine*, *49*(2-3), 190-193.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349.

Cummings, R. G., Elliott, S., Harrison, G. W., & Murphy, J. (1997). Are hypothetical referenda incentive compatible?. *Journal of Political Economy*, *105*(3), 609-621.

Dassen, F. C., Houben, K., & Jansen, A. (2015). Time orientation and eating behavior: Unhealthy eaters consider immediate consequences, while healthy eaters focus on future health. *Appetite*, *91*, 13-19.

Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology*, *25*(1), 1-17.

De Vries, J. H., Zock, P. L., Mensink, R. P., & Katan, M. B. (1994). Underestimation of energy intake by 3-d records compared with energy intake to maintain body weight in 269 nonobese adults. *The American Journal of Clinical Nutrition*, *60*(6), 855-860.

Dutch Nutrition Centre (2016) *Richtlijnen Schijf van Vijf*. Retrieved from http://www.voedingscentrum.nl/Assets/Uploads/voedingscentrum/Documents/Professionals/Schijf%20van%20Vijf/Voedingscentrum%20Richtlijnen%20Schijf%20van%20Vijf%202016%204.pdf

Freisling, H., van Bakel, M. M., Biessy, C., May, A. M., Byrnes, G., Norat, T., & Ocké, M. C. (2012). Dietary reporting errors on 24 h recalls and dietary questionnaires are associated with BMI across six European countries as evaluated with recovery biomarkers for protein and potassium intake. *British Journal of Nutrition*, *107*(6), 910-920.

Foster, G. D., Makris, A. P., & Bailer, B. A. (2005). Behavioral treatment of obesity–. *The American Journal of Clinical Nutrition*, *82*(1), 230S-235S.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349-360.

Gorzelitz, J., Peppard, P. E., Malecki, K., Gennuso, K., Nieto, F. J., & Cadmus-Bertram, L. (2018). Predictors of discordance in self-report versus device-measured physical activity measurement. *Annals of Epidemiology*, *28*(7), 427-431.

Haraldsdottir, J., & Sandström, B. (1994). Detection of underestimated energy intake in young adults. *International Journal of Epidemiology*, *23*(3), 577-582.

Health Council of the Netherlands. (2017a) *Dutch physical activity guidelines 2017.* Retrieved from https://www.gezondheidsraad.nl/sites/default/files/grpublication/beweegrichtlijnen2017_201708_0.pdf

Health Council of the Netherlands. (2017b) *Physical activity and risk of chronic diseases. Backgroud document to the Dutch physical activity guidelines 2017.* Retrieved from https://www.gezondheidsraad.nl/sites/default/files/grpublication/achtergronddocument_physical_activity_and_risk_of_chronic_diseases_0.pdf

Hebert, J. R., Clemow, L., Pbert, L., Ockene, I. S., & Ockene, J. K. (1995). Social desirability bias in dietary self-report may compromise the validity of dietary intake measures. *International Journal of Epidemiology*, *24*(2), 389-398.

Hebert, J. R., Hurley, T. G., Peterson, K. E., Resnicow, K., Thompson, F. E., Yaroch, A. L., ... & Nebeling, L. (2008). Social desirability trait influences on self-reported dietary measures among diverse participants in a multicenter multiple risk factor trial. *The Journal of Nutrition*, *138*(1), 226S-234S.

Heitmann, B. L., & Lissner, L. (1995). Dietary underreporting by obese individuals--is it specific or non-specific?. *BMJ*, *311*(7011), 986-989.

Higgs, S., & Thomas, J. (2016). Social influences on eating. *Current Opinion in Behavioral Sciences*, *9*, 1-6.

Hill, R. J., & Davies, P. S. W. (2001). The validity of self-reported energy intake as determined using the doubly labelled water technique. *British Journal of Nutrition*, *85*(4), 415-430.

Hoch, S. J. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. *Journal of Personality and Social Psychology*, *53*(2), 221.

Howie, P. J., Wang, Y., & Tsai, J. (2011). Predicting new product adoption using Bayesian truth serum. *Journal of Medical Marketing*, *11*(1), 6-16.

Howley, E. T. (2001). Type of activity: resistance, aerobic and leisure versus occupational physical activity. *Medicine & Science in Sports & Exercise*, *33*(6), S364-S369.

Johansson, G., Wikman, Å., Åhrén, A. M., Hallmans, G., & Johansson, I. (2001). Underreporting of energy intake in repeated 24-hour recalls related to gender, age, weight status, day of interview, educational level, reported food intake, smoking habits and area of living. *Public Health Nutrition*, *4*(4), 919-927.

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

Johnson, R. K., Goran, M. I., & Poehlman, E. T. (1994). Correlates of over-and underreporting of energy intake in healthy older men and women. *The American Journal of Clinical Nutrition*, *59*(6), 1286-1290.

Klesges, R. C., Klesges, L. M., Haddock, C. K., & Eck, L. H. (1992). A longitudinal analysis of the impact of dietary intake and physical activity on weight change in adults. *The American Journal of Clinical Nutrition*, *55*(4), 818-822.

Kong, A., Beresford, S. A., Imayama, I., Duggan, C., Alfano, C. M., Foster-Schubert, K. E., ... & Bain, C. E. (2012). Adoption of diet-related self-monitoring behaviors varies by race/ethnicity, education, and baseline binge eating score among overweight-to-obese postmenopausal women in a 12-month dietary weight loss intervention. *Nutrition Research*, *32*(4), 260-265.

Kristal, A. R., Shattuck, A. L., & Williams, A. E. (1992, June). Food frequency questionnaires for diet intervention research. In *Proceedings of the 17th National Nutrient Databank Conference* (pp. 110-125). International Life Sciences Institute, Baltimore, Md. Washington, DC.

Kuijer, R. G., & Boyce, J. A. (2014). Chocolate cake. Guilt or celebration? Associations with healthy eating attitudes, perceived behavioural control, intentions and weight-loss. *Appetite*, *74*, 48-54.

LaMonte, M. J., & Ainsworth, B. E. (2001). Quantifying energy expenditure and physical activity in the context of dose response. *Medicine and Science in Sports and Exercise*, *33*(6 Suppl), S370-8.

Laporte, R. E., Montoye, H. J., & Caspersen, C. J. (1985). Assessment of physical activity in epidemiologic research: problems and prospects. *Public Health Reports*, *100*(2), 131.

Leefstijlmonitor CBS. (2018, February 15). *Overgewicht volwassenen*. Retrieved April 1, 2018 from Volksgezondheidenzorg.info: https://www.volksgezondheidenzorg.info/onderwerp/overgewicht/cijfers-context/huidige-situatie

Lichtman, S. W., Pisarska, K., Berman, E. R., Pestone, M., Dowling, H., Offenbacher, E., Weisel, H., Heshka, S., Matthews, D. E., & Heymsfield, S. B. (1992). Discrepancy between self-reported and actual caloric intake and exercise in obese subjects. *The New England Journal of Medicine*, *327*, 1893–1898.

Loughran, T. A., Paternoster, R., & Thomas, K. J. (2014). Incentivizing responses to self-report questions in perceptual deterrence studies: An investigation of the validity of deterrence theory using Bayesian truth serum. *Journal of Quantitative Criminology*, *30*(4), 677-707.

Macdiarmid, J. I., & Blundell, J. E. (1997). Dietary under-reporting: what people say about recording their food intake. *European Journal of Clinical Nutrition*, *51*(3), 199.

Macdiarmid, J. I., & Blundell, J. E. (1998). Assessing dietary intake: who, what and why of under-reporting. *Nutrition Research Reviews*, *11*(2), 231-253.

Malhotra, N. K. (2006). Questionnaire design and scale development. *The handbook of Marketing Research: Uses, Misuses, and Future Advances*, 176-202.

Martin, C. K., Correa, J. B., Han, H., Allen, H. R., Rood, J. C., Champagne, C. M., ... & Bray, G. A. (2012). Validity of the Remote Food Photography Method (RFPM) for estimating energy and nutrient intake in near real-time. *Obesity*, *20*(4), 891-899.

Martin, C. K., Nicklas, T., Gunturk, B., Correa, J. B., Allen, H. R., & Champagne, C. (2014). Measuring food intake with digital photography. *Journal of Human Nutrition and Dietetics*, *27*, 72-81.

Menapace, L., & Raffaelli, R. (2015, May). Does Prelec Bayesian Truth Serum prevent Social Desirability Bias in Choice Experiments?. *Proceedings of the International Choice Modelling Conference 2015*.

Miller, S. R., Bailey, B. P., & Kirlik, A. (2014). Exploring the utility of Bayesian truth serum for assessing design knowledge. *Human–Computer Interaction*, *29*(5-6), 487-515.

Milton, K., Clemes, S., & Bull, F. (2013). Can a single question provide an accurate measure of physical activity?. *British Journal of Sports Medicine*, *47*(1), 44-48.

Muhlheim, L. S., Allison, D. B., Heshka, S., & Heymsfield, S. B. (1998). Do unsuccessful dieters intentionally underreport food intake?. *International Journal of Eating Disorders*, *24*(3), 259-266.

Novotny, J. A., Rumpler, W. V., Riddick, H., Hebert, J. R., Rhodes, D., Judd, J. T., ... & Briefel, R. (2003). Personality characteristics as predictors of underreporting of energy intake on 24-hour dietary recall interviews. *Journal of the American Dietetic Association*, *103*(9), 1146-1151.

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

Prentice, A. M., Black, A. E., Coward, W. A., Davies, H. L., Goldberg, G. R., Murgatroyd, P. R., ... & Whitehead, R. G. (1986). High levels of energy expenditure in obese women. *Br Med J (Clin Res Ed)*, *292*(6526), 983-987.

Prince, S. A., Adamo, K. B., Hamel, M. E., Hardt, J., Gorber, S. C., & Tremblay, M. (2008). A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *International Journal of Behavioral Nutrition and Physical Activity*, *5*(1), 56.

Robinson, E., Tobias, T., Shaw, L., Freeman, E., & Higgs, S. (2011). Social matching of food intake and the need for social acceptance. *Appetite*, *56*(3), 747-752.

Robinson, E., Thomas, J., Aveyard, P., & Higgs, S. (2014). What everyone else is eating: a systematic review and meta-analysis of the effect of informational eating norms on eating behavior. *Journal of the Academy of Nutrition and Dietetics*, *114*(3), 414-429.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2006). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. John Wiley & Sons.

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*(3), 279-301.

Rzewnicki, R., Auweele, Y. V., & De Bourdeaudhuij, I. (2003). Addressing overreporting on the International Physical Activity Questionnaire (IPAQ) telephone survey with a population sample. *Public Health Nutrition*, *6*(3), 299-305.

Samuel-Hodge, C. D., Fernandez, L. M., Henríquez-Roldán, C. F., Johnston, L. F., & Keyserling, T. C. (2004). A comparison of self-reported energy intake with total energy expenditure estimated by accelerometer and basal metabolic rate in African-American women with type 2 diabetes. *Diabetes Care*, *27*(3), 663-669.

Schechtman, K. B., Barzilai, B., Rost, K., & Fisher Jr, E. B. (1991). Measuring physical activity with a single question. *American Journal of Public Health*, *81*(6), 771-773.

Schoeller, D. A. (1990). How accurate is self-reported dietary energy intake?. *Nutrition Reviews*, *48*(10), 373-379.

Schoeller, D. A., & Van Santen, E. (1982). Measurement of energy expenditure in humans by doubly labeled water method. *Journal of Applied Physiology*, *53*(4), 955-959.

Shephard, R. J. (2003). Limits to the measurement of habitual physical activity by questionnaires. *British Journal of Sports Medicine*, *37*(3), 197-206.

Sims, J., Smith, F., Duffy, A., & Hilton, S. (1999). The vagaries of self-reports of physical activity: a problem revisited and addressed in a study of exercise promotion in the over 65s in general practice. *Family Practice*, *16*(2), 152-157.

The Netherlands Nutrition Centre. (n.d.) *Heb ik een gezond gewicht?* Retrieved May 14, 2018 from http://www.voedingscentrum.nl/nl/mijn-gewicht/heb-ik-een-gezond-gewicht.aspx%22%3Ehttp://www.voedingscentrum.nl/nl/mijn-gewicht/heb-ik-een-gezond-gewicht.aspx

Trijsburg, L., Geelen, A., Hollman, P. C., Hulshof, P. J., Feskens, E. J., van't Veer, P., ... & de Vries, J. H. (2017). BMI was found to be a consistent determinant related to misreporting of energy, protein and potassium intake using self-report and duplicate portion methods. *Public Health Nutrition*, *20*(4), 598-607.

Tsai, S. A., Lv, N., Xiao, L., & Ma, J. (2016). Gender differences in weight-related attitudes and behaviors among overweight and obese adults in the United States. *American Journal of Men's Health*, *10*(5), 389-398.

Van Beek, J., Antonides, G., & Handgraaf, M. J. (2013). Eat now, exercise later: The relation between consideration of immediate and future consequences and healthy behavior. *Personality and Individual Differences*, *54*(6), 785-791.

Van de Mortel, T. F. (2008). Faking it: social desirability response bias in self-report research. *The Australian Journal of Advanced Nursing*, *25*(4), 40.

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

Williamson, D. F., Serdula, M. K., Anda, R. F., Levy, A., & Byers, T. (1992). Weight loss attempts in adults: goals, duration, and rate of weight loss. *American Journal of Public Health*, *82*(9), 1251-1257.