Tell Me the Unverifiable Truth – Bayesian Truth Serum and Social Desirability Bias

An Experimental Analysis

A Master's Thesis

By

Marwan Abolmagd

M.Sc. Economics and Business

Specialization - Behavioral Economics

Erasmus School of Economics

Erasmus University Rotterdam, The Netherlands

Supervisor: Prof. Dr. Aurélien Baillon

Student Number: 431139ma

E-mail: 431139ma@student.eur.nl

Abstract

Self-report methods are commonly used for scientific research and other domains. However, the data gathered through these methods are subject to impairment and lack accuracy due to the influence of social desirability bias. This study tests if the Bayesian Truth Serum method would be an effective tool to decrease the prevalence of social desirability bias among questionnaire respondents, which is estimated to account for 10% to 75% of the variance in past literature. An electronic questionnaire was used to measure the respondents' vulnerability to social desirability bias using the Marlowe-Crowne Social Desirability Scale, and to measure the level of deception regarding six socially sensitive questions. Around half of respondents answered the six questions regularly and the other half was exposed to the Bayesian Truth Serum Method. The results showed that BTS significantly reduced the prevalence of social desirability bias among respondents. These findings indicate that the Bayesian truth serum method is a potentially suitable tool to counter the effect of social desirability bias.

Keywords: Social Desirability Bias, Bayesian Truth Serum, Subjective data validity, Self-report measures.

Table of Contents

1. Introduction

Subjective data acts as an essential source of information for scientific researchers and policy makers. However, the value of the data is limited by the quality of the source in terms of expertise and trustworthiness. Regarding expertise, measured and attained qualifications can be used to determine the degree of confidence that a given expert has the necessary competencies to make a trusted judgment. In terms of trustworthiness, given that there is no motive to be untruthful or there is objective data to compare it with, it would be fair to assume the subjective data is reliable. However, in many situations it is quite challenging to determine whether the respondent has an alternative motive and/or there is no objectively gathered data to compare the responses to. Furthermore, concerns also prevail when it comes to the means through which these data are gathered.

Self-report measures are often used to generate subjective data. However, various concerns about the validity of these measurements have been raised repeatedly. Different theoretical perspectives have addressed these factors that contribute to the validity issue. One is the cognitive perspective focusing on the underlying (cognitive) limitations that may arise from processes of recalling, comprehension, or any other (cognitive) operation required to complete the questionnaire. Another is the situational perspective, which is concerned with validity issues that may arise from external influences rather than internal. Both perspectives are not mutually exclusive; contrarily, theoretical frameworks that classify the validity issues may result in some overlapping and interconnectedness (Crandall, 1976). It is important to recognize that validity issues from both perspectives contribute to how truthful or how accurate responses are relative to reality. A respondent may consciously or unconsciously provide an untruthful answer. This possibly arises from carelessness, intentional deception, or is simply unintentional (Prelec and Weaver, 2014).

An important variable that could severely undermine the validity of a study that employs self-reporting is social desirability bias. In general, questionnaire participants have showed a pervasive tendency to present themselves in a socially desirable or favorable manner (Fisher, 1993). This is usually done in accordance with social norms. Participants tend respond more frequently in a positively perceived manner while less frequently choosing negatively perceived

answers, compared to their actual attitude or behavior. Thus, a number of social scientific studies may have been compromised.

For the aforementioned reasons, Prelec (2004) presented a scoring method for obtaining truthful subjective information which is called the Bayesian truth serum (BTS) method. He claimed that the method is based on an information scoring system that supposedly induces truthfulness from an expected value maximizing or Bayesian sample of respondents. This provides the researcher the ability to generate truthful reliable data even though he or she does not acquire an advanced working knowledge of the domain in question. Unlike other methods, the Bayesian truth serum method does not reward the most common responses and recognizes the answers with the minority views. Thus, respondents are not motivated to provide answers they believe are closer to the mean of the group.

In a study conducted by Weaver & Prelec (2013), it was pronounced that the effectiveness of BTS may mainly be driven by two main forces. The first is by decreasing the level of carelessness and intentional inattentive behavior while responding. Since the respondents are required to provide predictions regarding other respondents, the requirement of a deeper thought process should increase the level of engagement with the presented questions. Secondly, BTS can be used to provide an incentive system that rewards truth-telling, decreasing the overall level of intentional deception. It would be interesting to tackle a more specified form of intentional deception, aiming at acquiring social approval. Thus, this study is aiming to explore whether BTS method has a direct effect on social desirability bias, leading to the following research question:

Would the application of Bayesian truth serum method decrease the prevalence of social desirability bias?

The following section (2) provides the necessary theoretical background to derive the research question. Section 2 is divided into two main parts, the first thoroughly introduces and explains the BTS method, while the provide insights regarding the social desirability bias. Section 3 describes the experiment conducted to gather the necessary data for answering the research question. Subsequently, I provide the analysis and the results section including all the used data analysis tests conducted. The discussion section (5) then explains and highlights the outlined results from section 4, in addition to the limitations and further concerns related to the study. The conclusions drawn from the analysis are finally presented in section 6.

2. Literature Review

The Introduction section demonstrated the general motivation behind this research, more specifically how applying the Bayesian truth serum method would affect the prevalence of social desirability bias. The Literature Review section provides the necessary details of the fundamental components of the research question, namely Bayesian truth serum and social desirability bias. Additional details regarding Bayesian probability theory and performance based incentives are included to equip the reader with any necessary insights before tackling the hypotheses of the research study.

2.1. Bayesian Truth Serum

2.1.1. Bayesian Probability Theory

Bayesian probability theory was named after (and fundamentally developed) by the English statistician Thomas Bayes in the 1700s, initially mentioned in a paper titled *"An Essay towards solving a Problem in the Doctrine of Chances".* Using probability, the theory supplies a comprehensive framework to provide inference and reasoning. What is referred to as the "Bayesian Revolution" has been present for the last couple of decades, commonly employing the theory across many disciplines. Bayes theorem is mainly concerned with the manipulation of conditional probabilities, or the probability of occurrence of a chain of events (Olshausen, 2004). The joint probability of two events, A and B, can be expressed as

$$P(AB) = P(A \cap B)\, P(B)$$

$$= P(B \cap A)\, P(A)$$

Bayes rule can often be incorporated to interpret how beliefs are updated when new evidence emerges. A belief is normally updated when new information is acquired as time passes by. For illustration, try to think of a situation in which there is a specific hypothesis at hand. Meanwhile, as new evidence arises, the probability of accepting or rejecting this hypothesis varies.

Given the hypothesis is false, the obtained evidence may contradict it. If true, there is a certain probability the new evidence confirms (Angner, 2012).

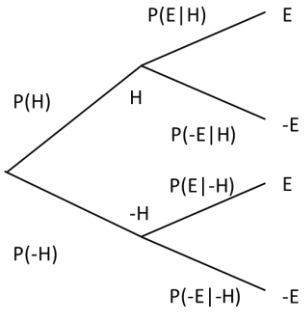


Figure 1: Bayesian Updating

Figure 1 gives structure to the given situation, where H represents the hypothesis being true and -H being false. The letter E represents evidence that supports the hypothesis being true, while -E represents the evidence supporting the hypothesis being false. The probability of H, referred to as the prior probability, is P(H) which indicates the probability that the hypothesis is true before knowing if the evidence confirms it. P(H|E) is the probability that the hypothesis is true given supporting evidence, which can be referred to as the posterior probability. The posterior probability can be expressed as:

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E|H) * P(H) + P(E|-H) * P(-H)}$$

This shows how beliefs are updated regarding H in the light of supporting Evidence E. The assigned probability to the hypothesis being true should change from P(H) to P(H|E).

2.1.2. What Is Bayesian Truth Serum?

Essentially, how the Bayesian truth serum (BTS) method works is instead of just asking the respondents to choose an answer that reflects their attitudes or behavior, they are asked to provide an empirical distribution of predictions of how the rest of the respondents will answer as well. The BTS method gives a high score to answers that are revealed to be more common than average prediction of the group of respondents. Thus, the information score is maximized with the answers that can be described as the most “surprisingly common”. As a result, any biases induced

by any consensus based method is removed, recognizing truthfulness in both minority and majority opinions.

Every question is assigned a particular score, which is determined by comparing the actual response with the average prediction of respondents. An example used in the initial paper written by Prelec (2004) proposed the following questions:

1. Do you think humanity will sustain past the year 2100?
    a. Definitely.
    b. Probably.
    c. Probably Not.
    d. Definitely Not.

2. In the past year, have you had more than 20 sexual partners?
    a. Yes.
    b. No.

3. Would you consider Picasso your favorite painter from the 20th century?
    a. Yes.
    b. No.

Each of the participants involved would choose a personal answer, in addition to a predicted distribution of how the rest of the participants would respond. Prior to that, the personal responses that are higher than average predictions of all respondents receive the highest information score and participants are compensated accordingly. For example, if 20% of the people answered yes to Picasso being their favorite 20th century artist (question 3) and the average predictions were 40% for this response, those who answered yes receive a positive information score for this answer. This is what is meant by a "surprisingly common" answer, indicating that the number of respondents who choose a specific answer was higher than collectively predicted. "The surprisingly common criterion exploits an overlooked implication of Bayesian reasoning about population frequencies. Namely, in most situations, one should expect that others will underestimate the true frequency of one's own opinion or personal characteristics" (Prelec, 2004).

Accordingly, it is logical to expect that a direct consequence of the typical Bayesian argument would be that the frequency of predictions of a specific opinion is at its highest if given from those who hold a similar view. Since having the opinion constitutes a reasonably valid signal regarding its popularity (Dawes, 1990).

2.1.3. The Assumptions of BTS

Bayesian Truth Serum relies on several assumptions that did not undergo proper practical testing (Weaver & Prelec, 2013). First, it models participants as Bayesian Agents when it comes to their predictions about the distribution of responses. The respondents consider personal beliefs as an informative signal about the distribution. They start with a generally common initial belief, followed by updating these beliefs with their own preference guidance. Considering this, Bayes' rule would strictly imply that those who pronounce the same preference will end up with the same predicted distribution. In addition, a state of Nash Equilibrium is created, all participants should assume other participants will behave rationally and provide a truthful response. An example that explains the logic behind the first assumption will be covered in a later section, indicating how updating beliefs would be considered rational. Another major assumption is that the sample used is large enough, that a single prediction or response would not cause any significant effect on the overall results. The level of skepticism regarding whether these assumptions will hold in an actual self-report setting should not be overwhelming. The idea of having an initial belief that changes by preference guidance is supported by psychological theory that will be discussed later. However, the distribution prediction by similar respondents differ greatly in practice (Weaver and Prelec, 2013). Regarding the second assumption, it is usually in the experimenter's control to gather a sample large enough that no single observation would influence the results significantly.

2.1.4. BTS and Performance Based Incentives

Behavioral studies often use incentives on research participants typically in exchange for their time and effort. However, such incentives have been debated in terms of controversy in effectiveness and inconsistency in application (Brase, 2009). For instance, participants are sometimes paid a flat-fee for showing up and participating in an experiment, this is most commonly used in experiments in the field of psychology. In contrast, in economics experiments,

the amount paid is based on the participant's performance on the given tasks. Another alternative is to combine both methods of payment, rewarding the participant with a baseline participation fee which can then be increased by how well he or she performed. Any of the discussed payment methods have their advantages and disadvantages, thus, there is no strictly dominant method that can be applied across all domains.

Choosing one method over the other should be a well examined decision since it can strongly affect the results. For instance, the idea of overly high stakes leading to performance anxiety is well documented in psychological research (Baumeister, 1984). Thus, providing high incentives that significantly vary based on performance is not necessarily optimal since it can lead to interference from a non-targeted variable (level of anxiety) on the final results. If the experiment for instance is trying to measure the relationship between IQ and performance in a specific task, induced anxiety due to high stakes would create major noise to the resulting data. This can be fixed by providing a flat fee incentive, since it would remove the cause of the increased level of anxiety with less distraction from the task. Ariely et al. (2008) explained this phenomenon by concluding that when stakes are high, the participant is distracted by the thoughts of performing badly and not being able to get a high reward. This divides the attention of the participants between the task at hand and the mental reflections about the outcome due to possible differences in performance.

Feasibility is also another major concern, as in some cases one method can provide a significantly better experimental setting but cannot be applied. An example would be an experiment that requires the participants to express their own views, beliefs, or opinions. There is no right or wrong answer, thus, there is no measure of performance to base the awarded amount on. This doesn't give the experimenter much of a choice but to go with a flat-fee incentive awarded to whoever finishes the questionnaire. The question here would be whether this is the optimal way to do it? What are the drawbacks in this setting? Can an alternative incentivizing method be applied to increase the reliability and the validity of the given answers?

In a typical self-report questionnaire example, a lot of things can go wrong to compromise the validity of the collected data. For example, a questionnaire can be too long, driving the respondent to get bored and not provide carefully thought through answers. There is

no tangible incentive to drive the participant to provide the extra effort to avoid any careless behavior while answering, the participant gets paid the same amount regardless of the answers. In this case, applying a task-related payment method is optimal but not feasible since there is no clear right or wrong answer. The Bayesian truth serum method can be used to solve this problem, assuming the truthful answer would be treated as the right one and performance is measured in terms of degree of truthfulness.

A study conducted by Weaver and Prelec (2013) has demonstrated the effectiveness of using the BTS method to provide performance based incentives. They conducted several experiments under which they tried different methods of questioning and measured the frequency of untruthful answers under each method. In three of the experiments they used a computer administered survey to ask the participants whether they recognize a shown item or not. Some of the items shown were foils, thus, if a participant answered that he or she recognizes any of these foils it is clear they are being untruthful since the item does not exist. Across the three experiments, the average percentage of recognized foils were significantly less when the questions were administered under the BTS than the control condition and similar truth-telling mechanisms.

Participants can be rewarded based on the information score (iScore). The iScore can be used to evaluate the answer, since there is no available answer key for each respondent to tell whether the answer is truthful or not (Prelec, 2004). The function for the information score for answer *k* would be

$$iScore = log\frac{\bar{x}_k}{\bar{y}_k}$$

Where $\bar{x}_k$ is answer's k actual relative frequency and $\bar{y}_k$ is the (geometric) average of k's predicted frequencies. One of the answers will then have a positive information score, and variance in the predictions is prone to decrease the value of $\bar{y}_k$ leading to an increase of the iScore.

Respondents can be indexed as $r\epsilon\{1,2,\dots\}$. Out of a set of m multiple choices, they then choose one answer t indicating whether the selected response k is the truthful answer (taking the value of

1) or not (taking the value of 0) of respondent r. The score for repondent r then combines the total iScore and the total prediction score.

$$rScore = \sum_k \bar{x}_k^r \log \frac{\bar{x}_k}{\bar{y}_k} + a \sum_k \bar{x}_k \, log \frac{y_k^r}{\bar{x}_k}, 0 \leq a$$

Where y_k^r is the prediction of answer k from respondent r, and $\bar{x}_k$ is the endorsement frequency of answer k. The prediction score on itself can be used as a penalty to how far did the prediction deviate from the actual distribution. If $y_k^r = \bar{x}_k$ will result with the best possible prediction score (0). Combining the information score and the prediction score gives us the function for the score of respondent r. The a is a constant aimed to provide a certain weight for the prediction score. In the current study $a = 0$, indicating that the participant is only rewarded based on the information score. A working example of how the prediction score is calculated is provided under Appendix A3.

2.1.5. Prior and Bayesian Updating

In this section, a step by step example will be used to explain how a Bayesian respondent would act in the situation paved by the BTS method. It shows how rational it is for a respondent to update his or her belief of distribution based on a respondent's own opinion. One of the questions presented to the participants in the current study is as follows:

Would you—for any reason—read your mate's email without his/her knowledge and permission?

a. Yes.
b. No

There are two categories of respondents, those who believe they would read their partner's email without their permission (Y category) and those who would not (N category). As mentioned earlier, using the BTS method, each participant is asked to provide their own response and a prediction of the distribution regarding how other respondents will answer the question. Using the assumption that respondents are pure Bayesians, we should expect that they all have a prior

distribution function regarding the prevalence of a response before answering the question themselves. We can refer to the prior estimation of Y as ω. For simplification, let us assume that the prior belief is either 0.6 or 0.85 who share their opinion. They also think that any of two distributions is equally likely. This can be expressed as ρ (ω = 0.6) = p (ω = 0.85) = 0.5. As Bayesian Agents, participants are expected to update their belief once they receive a signal or additional data. In this case, let us use the respondents answer as the signal to be incorporated with the prior distribution to come up with the posterior probability.

To clarify, we have two categories (possible responses) of people, those who will answer yes (Y) and those who will answer no (N). Focusing on yes category (Y), they have a prior belief that the percentage of respondents who share their belief is either 60% or 85%; and both are equally likely. We use t_r to express the category of the respondent r. What is the probability that a respondent with a belief ω= 0.6 or ω=0.85, will answer yes ($t_r = Y$)?

The answer is simply dependent on the initial distribution belief (ω), resulting with probabilities of 0.6 and 0.85 respectively. Assuming the participate answered yes, will the probability of having an initial belief of distribution of 0.6 or 0.85 (p=0.5) change? If yes, what are the posterior probabilities?

Prior belief about ω:

$$p(\omega = 0.6) = 0.5$$

$$p(\omega = 0.85) = 1 - \mathrm{P}\,(\omega = 0.6) = 0.5$$

E(ω) = 0.6 * 0.5 + 0.85 * 0.5 = 0.725

Posterior belief about ω:

$$p(\omega = 0.6|t_r = Y) = \frac{[p(t_r = Y|\omega = 0.6)_{*p(\omega=0.6)}]}{p(t_r = Y|\omega = 0.6)_{*p(\omega=0.6)+p(t_r = Y|\omega = 0.85)_{*p(\omega=0.85)}}}$$

$$p(\omega = 0.6|t_r = Y) = \frac{[0.6*0.5]}{0.6*0.5+0.85*0.5}$$

$$p(\omega = 0.6|t_r = Y) = \frac{0.3}{0.725} = 0.41$$

$$p(\omega = 0.85|t_r = Y) = 1 - 0.41 = 0.59$$

The calculation shows that once a respondent chooses one of the distributions that were equally likely, the probability regarding the beliefs gets updated than it initially was. This supports the idea that once a decision was made, perceived probabilities tend to change. Since the participant answered yes (Y), then the probability that 85% of the population answer yes increases and that of 60% decreases.

In parallel, let us ask the same question for responding no (Category N). What is the probability that a respondent with a belief ω= 0.6 or ω=0.85, will answer no ($t_r = N$)?

Again, depending on the initial distribution from the prior belief (ω), resulting with probabilities of 0.4 and 0.15 respectively. Assuming the participate answered No, will the probability of having an initial belief of distribution of Y equals to 0.6 or 0.85 (p=0.5) change? If yes, what are the posterior probabilities?

Prior belief about ω:

p(ω=0.6) = 0.5

P (ω = 0.85) = 1- P (ω = 0.6) = 0.5

E(ω) = 0.6 * 0.5 + 0.85 * 0.5 = 0.725

The posterior probabilities in this case will be as follows:

Posterior belief about ω:

$$p(\omega = 0.6|t_r = N) = \frac{[p(t_r = N|\omega = 0.6)*p(\omega=0.6)]}{p(t_r = N|\omega = 0.6)*p(\omega=0.6)+p(t_r = N|\omega = 0.85)*p(\omega=0.85)}$$

$$p(\omega = 0.6|t_r = N) = \frac{[0.4*0.5]}{0.4*0.5+0.15*0.5}$$

$$p(\omega = 0.6|t_r = N) = \frac{0.2}{0.275} = 0.73$$

$$p(\omega = 0.85|t_r = Y) = 1 - 0.73 = 0.27$$

After the change of probability regarding possible distributions, what is the expected distribution of those who will also answer yes if respondent r answered no ($t_r = N$)? and yes ($t_r = Y$)?

$$E(\omega|t_r = N) = p(\omega = 0.6) * p(\omega = 0.6|t_r = N) + p(\omega = 0.85) * p(\omega = 0.85|t_r = N)$$

$$E(\omega|t_r = N) = 0.6 * 0.73 + 0.85 * 0.27$$

$$E(\omega|t_r = N) = 0.67$$

Which is lower than E(ω) by 0.055, indicating that once the participant answered No his or her prior belief was updated to be equal to E(ω|t_r =N) = 0.67.

$$E(\omega|t_r = Y) = p(\omega = 0.6) * p(\omega = 0.6|B) + p(\omega = 0.85) * p(\omega = 0.85|t_r = Y)$$

$$E(\omega|t_r = Y) = 0.6 * 0.41 + 0.85 * 0.59$$

$$E(\omega|t_r = Y) = 0.75$$

Which is higher than E(ω) by 0.025, indicating that once the participant answered yes his or her prior belief was updated to be equal to E(ω|t_r =Y) = 0.75.

2.1.6. Supporting Behavioral Theories

<u>False Consensus Effect (Egoistic Attribution Bias)</u>

The social psychological theories that best supports the BTS assumptions is false consensus effect. It describes the phenomenon that social observers have the tendency to overestimate the relative commonness of their own beliefs and opinions. In past studies, the terms “egocentric attribution” or “attributive projection” were used to refer to this phenomenon (Heider, 1958; Holmes 1968). It appeared to be present when a correlation was found between subjects’ reported behavior and reported estimates about how their peers would behave (Murray, 1933). For example, another study revealed that when a student was asked how frequently s/he cheats and how frequently classmates cheat, the responses were positively correlated a significant amount of times

(Katz & Allport, 1931). This indicates that cheating students tend to believe that there are more cheating classmates relative to those who do not cheat.

Ross, Greene and House (1977) gathered evidence from four studies supporting what is referred to as "False Consensus Effect". In the first study, they presented 320 undergraduate students with brief hypothetical scenarios, each with two possible courses of action. The participants were first asked to provide an estimation of how other respondents would react to a given scenario. Upon completion of the estimations, participants were asked which course of action they would personally choose. The results showed that there was a general trend across all provided scenarios, that participants personally choose the alternative that is the most probable to be chosen by others. Indicating that they believed that the majority would act in a similar manner as they would. A second study was conducted to see whether next to actions people also have general tendencies to overestimate the amount of other people who share their preferences, habits, characteristics. The participants were handed a list of 34 personal description items and were asked to estimate the percentage of college students fitting each of the categories and to place themselves in categories of personal fit. The researchers hypothesized that participants who have placed themselves in a certain category would generally estimate the percentage of college students falling under that category to be larger relative to those who placed themselves in alternative categories. The results confirmed the hypothesized direction in 32 out of the 34 categories.

Selective Exposure

Another reason to why people overemphasize the commonality of their own beliefs, is that they simply notice more those who do. People tend to associate with and know others who share their interests, background, experiences, and values. This implies that people selectively expose themselves disproportionately to those similar individuals, relative to someone else who has as different outlook. In addition, those similar individuals tend to respond in a similar manner in various circumstances. This idea is referred to as selective exposure effect (Ross, Greene, and House; 1977). The idea of selective exposure can be supported further by two major psychological concepts, cognitive dissonance and confirmation bias. Cognitive dissonance explains that when people have contradicting beliefs or attitudes, it results with an uncomfortable state referred to as dissonance which makes them unhappy; they then take necessary action to make this dissonance

go away (Killian, Festinge, Riecken, & Schachter, 1957). Selectively noticing and being around those who have similar beliefs and attitudes can be viewed as way to deal with dissonance that might arise from being exposed to the contrary. Confirmation bias on the other hand explains that when people have a specific theory or belief in mind, their brain automatically notices information that supports this theory and neglect contradictory information (Nickerson, 1998). This might explain why noticing and recalling examples of those who share an individual's belief is less effortful than those who do not.

2.1.7. Contradicting Behavioral Theory

Pluralistic Ignorance

A social situation in which pluralistic ignorance takes place is if a group of individuals share an attitude towards a specific proposition, yet all act contrary to it, and all of them believe that all the other individuals in the group have a conflicting attitude to this proposition (Bjerring, Hansen, & Pedersen, 2014). An example for this is a study conducted by Katz and Allport (1931) with a group of students which found that students individually did not have any objections to minorities joining the fraternity, but each believed that the other students would object. As mentioned earlier, the BTS method assumes that people overestimate the prevalence of their own opinion, which is contradicting to what pluralistic ignorance is pointing to. BTS would suggest that if someone holds a specific proposition, this individual should assume that this proposition is more widely shared than it actually is, not the other way around.

Prentice & Miller (1993) suggested two reasons that would explain the phenomenon of pluralistic ignorance. They referred to the first as differential interpretation hypothesis. It indicates that even though an individual would act misleadingly in agreement with a specific proposition that he or she disagreed with, this individual fails to interpret that others are misleading as well. The second is differential encoding hypothesis, which indicates that misleading individuals fail to realize to what extent their behavior appear in favor of the proposition they secretly are against. Two more reasons that were covered by complementary research are the attempt to maintain social identity and minority influence. Maintaining social identity would be considered a source of pluralistic ignorance since it is basically synchronized with the definition of the phenomenon. Pluralistic ignorance can be considered an unexpressed standpoint that was not endorsed as an attempt to coordinate and maintain affiliation with the social group. Minority influence refer to the

situation when the majority of the group takes the point of view of a minority to reflect the norms and identity of the whole group (Halbesleben & Bowler, 2007). The four given reasons are examples where the BTS assumption regarding using one's opinion as a signal of the its distribution clearly fails to hold. In the case of pluralistic ignorance, the misleading behavior of other individuals in the group is the reference point of distribution.

2.2. Social Desirability Bias

2.2.1 What is social desirability Bias?

The idea of social desirability bias (SDB) was first introduced by Allen Edwards in 1957. He described the concept as the tendency of respondents to answer questions in a manner that is socially acceptable, aiming to acquire approval of others. In his research, he investigates how SDB was plaguing research in personality through the distortion of what he referred to as the "Lying Factor" in responses. This "Lying Factor" is prone to be evoked by three main factors; the testing or experimental setting, the respondent's motives, the respondent's expectations about the possible evaluative consequences of their behavior or responses" (Edwards, Diers, & Walker, 1962).

Researchers employing questionnaires and other self-report measures rely on participants responding truthfully to come up with meaningful conclusions. However, participants may modify the responses as an attempt to deceive themselves, confirm with socially acceptable concepts, avoid being criticized, or acquire social approval (Huang et al., 1998). The more socially sensitive the question is, the higher the likelihood the participant would provide a socially desirable response (King & Brunner, 2000). A study was conducted using both self-report measures and activity trackers in an attempt to determine the level of physical activity of the participants. The study also included a social desirability questionnaire, which will be discussed in more details later in this section, and found that those with a high SD score significantly overestimated the level of physical activity performed (Adams et al., 2005). The effect of social desirability bias can be crucial, as one study estimated that 10% - 75% of variance in participant responses is due to SDB (Nederhof, 1985). This high level of variance due to SDB can easily obscure the results of the relationship between the primary variables under investigation, and compromise the validity by deviating the results from what actually occurs in the real world (King & Brunner, 2000).

2.2.2. Measuring Social Desirability Bias

Various social desirability (SD) scales were developed; however, the one that was most widely accepted and used is the Marlowe-Crowne Social Desirability Scale (MCSDS). Recently, the Balanced Inventory of Desirable Responding (BIDR) is used more often (Lambert, Arbuckle, & Holden, 2016). Normally, social desirability scales get administered to aid the establishment of discriminate validity of the primary test employed. Usually a higher inter-correlation between the SD score and the target test indicates that the answers are confounding socially desirable responses, and the opposite with a lower one (Paulhus, 1991). In the case of MCSDS, higher SD score indicates that the respondent is likely to provide a socially desirable response on any socially sensitive content. A regression analysis can be conducted to determine to what extent the variance in responses is attributed to social desirability bias. Further, a number of operations maybe conducted to take corrective action and eventually increase the validity of the test.

The MCSDS was introduced by Douglas Crowne and David Marlowe in 1960. They concluded that a considerable number of individuals tend to portray themselves in a relatively positive manner by exaggerating their positive attributes, inflating their strengths, diminishing their deficiencies, and trivializing their failures. Crowne and Marlowe identified a set of behaviors related to social desirability bias, and used existing personality inventories to try and extract them. The items included initially were selected by following a set of criteria concerned with cultural approval, and had minimal pathological implications. These criteria were set to make sure that the items identified are positively perceived across cultures, and is more concerned with the general populations and not for diagnosis of abnormal behavior. Fifty items met the initial criteria. The list of items was then submitted to graduate students and faculty members of the psychology department of Ohio State University to be rated according to social desirability. Out of all people, 90% agreed on 47 out of the 50 listed items. The 47 items were then given to 76 undergraduates students and they were asked whether they engage in these behaviors. Out of the 47, only 33 items were significantly related to the aggregate total. The result is a list of 33 items that can be used to differentiate between those who are more likely to exhibit social desirability bias and those who are not. The 33 items were used to formulate the final form of the scale, had a calculated internal consistency coefficient of 0.88. Additionally, the obtained test-retest correlation was 0.89 (Crowne & Marlowe, 1960).

The Balanced Inventory of Desirable Responding (BIDR) is another social desirability scale that provides measurements based on two constructs, self-deceptive enhancement and impression management. Self-Deceptive Enhancement (SDE) can be described as the tendency to unconsciously provide unrealistic self-reports that tend to be honest but positively biased. On the other hand, Impression Management (IM) refers to the tendency to consciously provide inaccurate self-descriptions (Lambert, Arbuckle, & Holden, 2016). BIDR contains 40 items was proven to have adequate reliability by having an internal consistency coefficient alpha of 0.83, and a test-retest correlation of 0.67. Concurrent validity as a measure of social desirability bias is shown by a correlation of scores of 0.71 with the MCSDS (Paulhus, 1984).

Lambert, Arbuckle, & Holden (2016) conducted a number of studies to compare the efficacy of MCSDS and BIDR in detecting deceivers. They asked all participants to answer the MCSDS and both parts of the BIDR (SDE and IM) as they were undergoing a military induction screening. In two out of the three studies, each of the participants was allocated to one of three conditions. The first condition asked the respondents to complete the questionnaires as instructed, they were asked to fake their answers in an attempt to maximize their chances to be inducted in the second condition, and to minimize their chances for the third. The participants were informed that there will be a validity check to detect faked responses, and those who are least detected receive a monetary reward. They were then asked to report the extent to which they complied with their condition-specific instructions. For study three the scenario differs, as they were being screened for a government job that requires handling confidential material and money. In addition, they were informed about the validity checks but were not incentivized for their efforts to evade detection. The study hypothesized that the IM part of the BIDR will be a more accurate detector of faking respondents than the MCSDS. This hypothesis was rejected as the MCSDS outperformed the BIDR consistently, and the study concluded that MCSDS should be reconsidered as the gold standard for measuring SDB. Due to these results and conclusion, and the fact that MCSDS is more reliable (with a higher test-retest correlation of 0.22 and internal consistency coefficient of 0.05) The MCSDS will be employed in the current study as the measure of social desirability bias.

2.2.3. Ways to reduce Social Desirability Bias

Several studies have shown that SDB occurs due to two main reasons. First, it happens as respondents attempt to manage their impressions and look better to the researcher. Second, it occurs due to self-perception enhancement, to look better to themselves. One claimed approach to decrease SDB is to assure the respondents absolute anonymity and willingness to participate in the survey. The logic behind it is that increased assurance of confidentiality should make the participants believe that there is no need to change the responses to be more socially acceptable. An experiment was conducted at the University of Mannheim in 1988 to investigate the relationship between questionnaire confidentiality and SDB. Participants were handed a questionnaire that varied in instructions between three different levels of confidentiality assurance. The results showed that there was a decreasing willingness to participate as the level of confidentiality increased. Although the study is not directly related to SDB, the results show that increased confidentiality assurance elevated the participants' social desirability guard. They had an increased expectancy that they would be asked uncomfortable questions, ultimately decreasing the participation rate. If the participants are already evading participation due to threatening questions, the impact of social desirability bias may be stronger with those who respond. In other words, increased assurance of confidentiality may backfire and increase the effect of social desirability bias (Paulhus, 1988).

Another method introduced by Raghubir and Menon (1996) to reduce SDB is counter-biasing. The method involves the introduction of what is normally considered as an undesirable behavior as more socially acceptable. They conducted a study in which they had an experimental and a control group. The experimental group was initially exposed to counter-biasing information before the questionnaire and the control group was not. The information the experimental group was exposed to was that a typical student called John has protected sexual intercourse once every five times. The results showed that the reported likelihood of using protection during sexual intercourse was significantly less when the participant was exposed to counter-biasing information.

The third commonly used method to decrease the effect of SDB was studied by Fisher in 1993, particularly focusing on indirect questioning. He hypothesized that indirect questioning should decrease the distortion created by SDB when participants were asked about private opinions. The method basically relied on asking the participant to reflect on the nature of the

external world, rather than him or herself. The idea is that their attitudes and behavior towards the subject matter would be unconsciously projected. He conducted a study in which he had four different conditions; anonymous direct questioning, non-anonymous direct questioning, anonymous indirect questioning, and non-anonymous indirect questioning. The results showed that all students made similar evaluations regarding socially neutral questions, but significantly reported more socially desirable answers when asked directly. There was no significant difference observed between the anonymous and not anonymous groups. The results indicate that indirect questioning can be an effective method to reduce SDB, but not the same case for increased anonymity.

The BTS method works by requiring the participant to provide a response and the expected distribution of responses among other respondents. In a way, it is asking the respondents to reflect on the external world besides themselves. It is fair to say then that it incorporates direct and indirect questioning. Since indirect questioning is partially involved, it is expected that BTS will reduce the prevalence of the effect of SDB. In addition, providing truth-telling incentives is expected to counter SDB. Hence, the following hypothesis was developed and the study was formed accordingly.

Hypothesis: Bayesian Truth Serum method decreases the effect of social desirability bias in self-report measures.

3. The Experiment

The experiment Consists of 2 between-subject conditions. It was carried out via an electronic survey through social media means. The majority of the participants were from the Middle-East and the Eurozone. About half of the participants were male and half were female mostly around the age of 24. All 54 participants received an anonymous link to the questionnaire via Facebook. The questionnaire was conducted through Qualtrics, and differed per treatment and was distributed using even randomization options.

3.1. The Questionnaire

The questionnaire was constructed out of two main parts. Starting with the 33 questions directly taken from the Marlowe-Crowne Social Desirability Scale (MCSDS) to determine the degree of influence of social desirability bias on each participant. The second part contained six mildly to moderately socially sensitive questions presented in two different ways according to the condition that each participant was allocated to randomly. For ease and clarity, I will refer to the second part as the Social Sensitivity Questionnaire (SSQ). Randomization was done through an automated process that allocated each participant to one of the conditions based on a 50% probability. In the control group, the participant was asked the questions directly. As for the treatment group, the BTS method was applied, through which the participant was asked to answer the question and provide a distribution prediction of how 100 other participants would respond.

All participants were instructed to respond to section one as follows:

> "This section contains 33 statements concerning personal attitudes and traits. Read each item and decide whether the statement is true or false as it pertains to you personally".

All questions were identically copied from the MCSDS questionnaire with the same order. As previously mentioned, based on the response for each statement, the participant may or may not receive a point for each of the statements based on the response. Thus the maximum social desirability score the participant may receive is 33, while the minimum is 0. The higher the score, the higher the likelihood the participant would endorse socially desirable responses when it comes

to other socially sensitive questions. In other words, those who have a higher MCSDS score fall into social desirability bias more often. All MCSDS questions require the participant to answer by selecting "True" or "False". To provide some examples, the first three questions were presented as follows:

Q1. Before voting I thoroughly investigate the qualifications of all the candidates.

- True.
- False.

Q2. I never hesitate to go out of my way to help someone in trouble.

- True.
- False.

Q3. It is sometimes hard for me to go on with my work if I am not encouraged

- True.
- False.

The full questionnaire is available under Appendix A.1. Depending on the answer, the participant either receives a point or not for each question. The more socially desirable response for each question receives the point. For the first three questions, if a participant answered "True" to all of the presented questions; his or her MCSDS score would be two. Given that the socially desirable answer for the first two questions is "True", and "False" for question number three.

The second part of the questionnaire differed according to the condition. In the No-BTS condition (control group), the participants were simply asked to answer the six questions. Regarding the BTS condition (treatment group) it was more complicated since the respondent was requested to provide an answer, prediction of others answers, and read an explanation on how BTS work. The following is an example of one of the SSQ questions, as presented to No-BTS versus the BTS conditions:

No-BTS Condition

Q37. Would you—for any reason—read your mate's email without his/her knowledge and permission?

- Yes.
- No.

BTS Condition

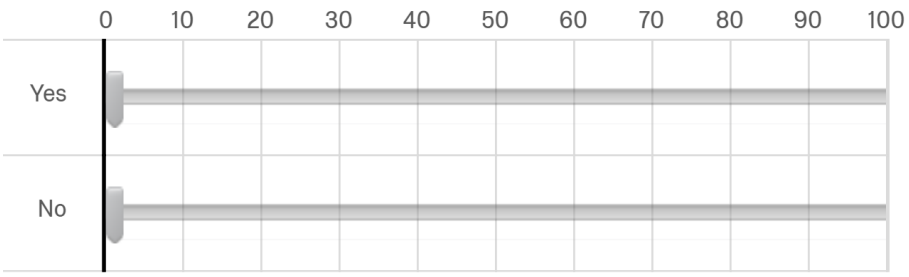
Q37a. Would you—for any reason—read your mate's email without his/her knowledge and permission?

- Yes.
- No.

Q37b. Out of 100 respondents, how many do you think will answer "No" and

how many will answer "Yes" to the same question?

Make sure Yes + No = 100



The instructions were given to the BTS condition participants as follows:

> Answer the following 6 two-part questions as indicated.
>
> For each question you answer you will receive an "information score". The sum of all information scores will then be calculated to come with your overall "Truth Score."
>
> Truth scoring, recently invented by an MIT professor and published in the academic journal Science, rewards you for answering truthfully. Even though only you know if you really answered truthfully or not, people who tell the truth score higher overall.
>
> You are most likely to maximize your potential donations if you answer every item truthfully. By "truthfully," we mean: consider each question carefully, answer honestly, and take care to avoid mistakes.

Prior to answering the social sensitivity questionnaire, all participants were asked to provide their age, gender, nationality and education level.

3.2. The Incentives

As mentioned earlier, the BTS method can provide adequate means to design an effective incentive system. This is based on the information score the participant receives for each question, providing the possibility to provide a performance incentive system that rewards truthfulness. In the meantime, since the questionnaire is mainly based on socially sensitive questions, compromising the anonymity was a major concern as it may lead to more deceiving responses from the participants. However, since it was mentioned in earlier literature that direct anonymity assurance may back fire as it may scare participants of by giving them the impression that there will be embarrassing questions; anonymity was not explicitly assured but the expiremental design pronounced it by not asking for contact details. To overcome the anonymity concern, instead of paying the participants directly which would have required asking for contact information, the amount each participant gained was donated on their behalf to a charity project of their choice.

Another limitation regarding the incentives was the available funding dedicated to the study. To try and minimize the effect of this limitation a couple of techniques were used.

3.2.1. Currency of Donation

The first would be expressing the reward in a more numerous currency, making use of the money illusion and numerosity heuristic. Money illusion refer to a phenomenon observed by Shafir, Diamond and Tversky (1997), through which they explained how people rely more on the nominal or face value to evaluate a transaction rather than real value. Another study was conducted to see how people deal with foreign currencies differently. The results showed that when a foreign currency has a high nominal value than a home currency, people are less likely to spend if they use the foreign currency. In the case when the foreign currency has a lower nominal value then the home currency, they are more likely to experience increased spending (Raghubir & Srivastava, 2002). This can be demonstrated by a study conducted when the EU countries started adapting the euro currency, and consumer price estimation at the time in different countries. The study used the Italian currency which was at an exchange rate of 1,936.27 liras per euro, and the Irish currency which was at an exchange rate of 0.788 Irish Pounds per euro. People from both countries were asked then to estimate the price of some consumer goods in euros. The hypothesis was that participants from Italy will generally overestimate the value of the goods, and participants from

Ireland will slightly underestimate it. The results showed that Italian participants overestimated all 12 products by an amount ranging from 9% to 70%, while Irish participants underestimated all prices by an amount ranging from 1% to 21% (Del Missier, Bonini, & Ranyard, 2007).

Numerosity heuristic was reported as the tendency to judge a stimuli based on the number of units. A study was conducted by exposing participants to one whole circle and a circle divided into nine pieces, followed by asking them to estimate the surface area of each circle. The results showed that more participants perceived the divided circle larger than the fully intact one. The interpretation was that people tend to rely on the notion of numerosity to provide an estimation of the perceived amount, especially when their cognitive resources are drained (Pelham, Sumarta, & Myaskovsky, 1994). These concepts were used in this study to provide an illusionary inflated value compared to the real value of the amount. Each participant was awarded a maximum of 20 Egyptian Pounds, which was equivalent to roughly 2 euros at the time.

3.2.2. Active Choice

The second technique used to increase the perceived value and effectiveness of the incentive, was giving the participant the option to choose one of three charities to donate the money to. This should help maximize the effect of the incentive through two different means. The first is very intuitive; by providing more than one option, there is a higher probability the participant would find an option related to an intimate cause that he or she might empathize with. This should increase the intrinsic motive to try to answer as truthfully as possible to maximize the amount of money to be donated.

The second is more related to complex behavioral theory. The idea is that making a choice should increase the level of commitment relative to just informing them that the money will go to one specific organization. A study conducted by Stutzer, Goettea and Zehnder in 2011 investigated if active choice increases the willingness to contribute and encourage prosocial behavior. They hypothesized that “issue-specific altruistic preferences” and increased commitment would be formed as a result of confronting people with a choice rather than being defaulted. A field experiment was conducted using blood donation, in one condition people were invited to donate and in the other they were asked to choose whether they would like to donate or not. The results

revealed that actively choosing to donate significantly increased the actual donations relative to just being invited, even if there was no preferences formed beforehand.

Extending the findings to the current conditions, it is fair to assume that participants would feel more committed to try and maximize the amount to be donated, ultimately leading to more effective monetary incentive. Each participant was given three charities to choose from, all of which are located in Egypt. The choice was presented to the BTS condition as follows:

> As mentioned earlier, truthful answers are the best way to maximize the "Truth Score." This will then determine the amount of money to be donated on your behalf to an Egyptian charity organization or project of your choice. Just by participating, EGP 5 (Egyptian Pounds) will be donated. This amount can increase to up to EGP 20 based on your "Truth Score." In other words, each truthful answer will contribute with a maximum of EGP 2.50.
>
> Based on your "Truth Score", you will gain a certain amount of money that will then be transferred to a charity organization under the project that you choose. The Available projects are as follows, please select your preferred project:

1. [Kiosk Project (sustainable income)] It aims to provide a poor family with a source of a steady income, in this case the source is a Kiosk fully equipped, licensed, and contains initial goods valued at EGP 3000. The establishment of each kiosk costs the organization EGP 8000 and is estimated to provide a monthly income of EGP 500.

2. [Winter Blankets Project] Aims to help the most vulnerable families in villages and communities to face the winter cold by providing blankets during the coming winter season. Each blanket costs around 30 EGP.

3. [The Children's Cancer Hospital Foundation (CCHF) 57357] The CCHF 57357 is a legal independent non-profit organization with a vision "to be the unique worldwide icon of change towards a cancer-free childhood."

Mentioning the maximum value contributed by an individual question was meant to emphasize how one single question can influence the end amount significantly. The No-BTS condition was presented with identical choices, but the instructions differed. They were informed that just by completing the second section, 20 Egyptian Pounds will be donated to the charity they choose.

3.2.3. Impact of Projects and Donations

In an attempt to increase the perceived value and impact of the amount to be donated, two main selection criteria were considered. First, the impact and the importance of the project. In this

case, donating to help children with cancer might seem more of a priority than keeping people warm. Based on this reasoning the impact of donating to the children cancer hospital is labeled "High" and "Low "for winter blankets. The second criteria would be the impact of one full single donation, which is again labeled "High", "Medium", or "Low". The rationale behind it is that the value of the donation is also influenced by how attainable the desired goal is. For instance, a participant might choose winter blankets because her efforts are more pronounced in the sense that if she answers all questions truthfully, she buys a person in need two thirds of a blanket. A single contribution seems to be quite valuable relative to the kiosk project for instance, which requires another 399 full contributions to be attained as shown in table 1.

Table 1: Used Criteria for Choosing the Presented Charities

Project	Impact of the project	Cost of a unit	Single Donation	Impact of a single Donation	Number of full donations to cover a single unit
Kiosk Project	Medium	8000	20	Medium	400
Winter Blankets	low	30	20	High	1.5
The Children's Cancer Hospital Foundation	High	very high & unspecified	20	Low	very high & unspecified

A study was done by Dan Ariely (2008), presenting participants with one of two scenarios. In the first one, he asked if they were considering to buy a $25 pen and would then hear that another shop 15 minutes away sells the same pen for $7 less, would they go to the other store. The other scenario was identical but the object was a suit valued at $455 while the other shop also sells it for $7 less. The difference in amount of responses between both scenarios was significant: 73% said they would go the other store for the pen but only 21% for the suit. He later explains that the difference is due to the relative proportion of the discounted amount not its absolute value. In other words, $7 was perceived of a much greater value when the initial amount was $25 rather than $455. This conclusion can be extended to the relative value of the donated amount, which would be perceived at its highest in the winter blankets charity condition.

Both criteria were chosen so that each project would have an advantage. It is expected that for some it would be more motivating if the project is important, and others might care more about

how pronounced their contribution would be. The Kiosk project choice has more of a balanced equation, in addition to the advantage of providing something more sustainable.

4. Data, Research and Analysis

4.1.Description

The study contained 56 independent observations on an individual level, and one observation at the session level. Two individual observations were later excluded due to incomplete responses. Through means of randomization, 26 (48%) participants underwent treatment and were in the BTS condition; while the other 28 (52%) participants were in the No-BTS condition.

The sample exhibited a greater majority of 56% (30) male and 44% (24) female participants, all of which aged between 21 and 35 with an average age of 24.89. Participants were from various regions including 22 from the Middle East, 27 from Europe, 4 from Asia, and 1 from South America. Regarding their education level, 30 participants worked on or held an undergraduate Bachelor's degree and 24 had a Master's degree. Table 2 clearly mentions and describes the variables used in the study.

Table 2: Included Variables & Description

Variable	**Type**	**Description**	**Range**
Social Desirability Score(SDS)	Interval Scale	Indicates the participant's score on the Marlow-Crowne Social Desirability Scale.	0-33
Condition	Binary	Indicates if the participants answered the questionnaire under the BTS condition (1) or not (0).	0-1
Racial Involvement Mild	Binary	indicates if a socially desirable response was given for question 6.	0-1
Racial Involvement Moderate	Binary	indicates if a socially desirable response was given for question 1.	0-1
Online Dating	Binary	indicates if a socially desirable response was given for question 2.	0-1
Empathy	Binary	indicates if a socially desirable response was given for question 3.	0-1

Privacy Respect	Binary	indicates if a socially desirable response was given for question 4.	0-1
Drug Use	Binary	indicates if a socially desirable response was given for question 5.	0-1
Socially Sensitivity Questionnaire Score (SSQS)	Interval Scale	Indicates out of the 6 questions, how many socially desirable responses were given.	0-6
Adjusted SSQS	Interval Scale	SSQS excluding the Racial Involvement Mild variable	0-5
Age	Continues	indicates age of the participant.	21-35
Gender	Nominal	2 categories that indicates if participant is Female (1) or Male (2).	1-2
Nationality	Nominal	15 categories that indicates nationality of the participant: Egyptian (1) Dutch (2) Indian (3) Bengali (4) German (5) Saudi Arabian (6) Slovenian (7) Polish (8) Greek (9) Ukrainian (10) Palestinian (11) Brazilian (12) Italian (13) Austrian (14) Finnish (15).	1-15
Region	Nominal	4 categories that indicates region of participant: Middle Eastern (1) European (2) Asian (3) South American (4).	1-4
Education Level	Ordinal	2 categories that indicates Education Level: Bachelors (1) Masters (2).	1-2

The questions used to express the variables 3 to 8 are as follows. The answers highlighted in bold represent the ones that was considered the socially desirable ones.

Racial Involvement Mild

Would you prefer to go out with someone of your own skin color / racial background?

- Yes.
- **No.**

Racial Involvement Moderate

If you were going to have a child, would it be a problem if the other parent wouldn't be of the same ethnicity as you?

- Yes.
- **No.**

Online Dating

Would you be willing to meet someone through an online dating app or site?

- Yes.
- **No.**

Empathy

Which is worse: starving children or abused animals?

- **Starving Children.**
- Abused Animals.

Privacy Respect

Would you—for any reason—read your mate's email without his/her knowledge and permission?

- Yes.
- **No.**

Drug Use

Have you used psychedelic drugs (LSD, mescaline, peyote, etc.) ?

- Yes.
- **No.**

4.2. Research and Analysis

This section mentions and describes which tests were used to conduct the analysis and the results in attempt to tackle the following hypothesis.

Bayesian Truth Serum application will decrease the difference of social sensitivity questionnaire score between individuals who score high on the Marlow-Crowne Social Desirability Scale questionnaire (have a high social desirability score) and those who score low. Thus, decreasing the effect of Social Desirability Bias.

Expressed as:

H0. Bayesian Truth Serum does not decrease the correlation between Marlow-Crowne Social Desirability Scale and the Socially Sensitivity Questionnaire Score.

H1. Bayesian Truth Serum decreases the correlation between Marlow-Crowne Social Desirability Scale and the Socially Sensitivity Questionnaire Score.

An additional measure was considered to test one of the fundamental BTS assumptions on the given data.

4.2.1. Testing the BTS Assumption

In this section Table 3 is presented to test one of the major BTS assumptions. The assumption that those who provide a certain response, will give a higher prediction of the prevalence of their

answer relative to those who choose the other alternative. The first column indicates the question number. The second shows the possible responses for each of the questions, referring to the first as A and the second as B. In the following two columns, the average predictions for A and B responses by those who answered A or B are provided. For example, as shown in table 3, those who responded "yes" for the first question predicted that 76% of other respondents will answer "yes". On the other hand, those who answered "no" only predicted that 49% will answer "yes". The difference between the prediction of the prevalence of A responses by A respondents and the prediction of prevalence of A responses by B respondents is in this case 27%; which is provided in the fifth column. A positive value in the fifth column indicates that the BTS assumption that predictions of A% from A respondents > predictions of A% from B respondents. In the current study, this is the case for all six questions.

Table 3: Includes the results of testing the BTS Assumption for each of the questions.

Question	*Responded*	*Average Prediction of A %*	*Average prediction of A% from A respondents - average prediction of A% from B respondents*
1	A – Yes	76	27
	B – No	49	
2	A – Yes	53	3
	B – No	49	
3	A - Starving Children	76	7
	B - Abused Animals	68	
4	A – Yes	41	6
	B – No	35	
5	A – Yes	43	4
	B – No	39	
6	A – Yes	65	23
	B – No	43	

4.2.2. Descriptive Statistics

This section is meant to provide basic descriptive statistics to increase familiarity with the data and its trends between both conditions. Table 4 presents the frequency of socially desirable responses for each of the six questions in the social sensitivity questionnaire. The frequency of socially desirable responses dropped when applying the BTS method in three of the six questions, most noticeably under the privacy respect, racial involvement moderate questions with a difference of 13.7% and 16% respectively. Racial involvement mild questions only dropped by 6%. Drug use increased by 9.9%, and empathy by 2.2%. Overall in both conditions, the claimed socially desirable responses seems be provided by the majority. Since socially desirable responses are expected to be more prominent, this high level of prevalence can indicate that on a group level the responses are in fact socially desirable in five out of the six questions. Regarding the online dating question, this trend is strictly reversed, with a prevalence of 17.9% under the No-BTS condition and 46.2% under the BTS condition. This may indicate that the majority of the observation did not share the belief that responding "no" is socially desirable, but contrary. This possibility is taken into account, Appendix B2 contains the results of a supplementary ordered logistic regression of a modified SSQS variable that considers responding yes to the online dating question as the socially desirable response.

Table 4: Descriptive Statistics of SSQ Responses

	No-BTS		BTS	
Variable	Frequency of Socially Desirable Response	% of Socially Desirable Response	Frequency of Socially Desirable Response	% of Socially Desirable Response
Racial Involvement Mild (No)	20	71.4%	17	65.4%
Racial Involvement Moderate (No)	26	92.9%	20	76.9%
Online Dating (No)	5	17.9%	12	46.2%
Empathy (Starving Children)	22	78.6%	21	80.8%
Privacy Respect (No)	20	71.4%	15	57.7%
Drug Use (No)	22	78.6%	23	88.5%
Observation	28		26	

In Table 5, the mean and standard deviation of the variables SDS, SSQS, adjusted SSQS, and age are presented for both conditions. On average, the BTS condition group scored higher on the SDS by 1.4 points relative to the No-BTS group, but both groups exhibit a similar standard deviation. The mean and standard deviation for the SSQS and the adjusted SSQS are very similar for both condition groups.

Table 5: Descriptive Statistics of Interval and Continues Variables

	No-BTS		BTS	
Variable	Mean	Standard Deviation	Mean	Standard Deviation
Social Desirability Score (SDS)	16.9	4.4	18.3	4.1
Social Sensitivity Questionnaire Score (SSQS)	4.1	0.9	4	0.8
Adjusted SSQS	3.4	0.7	3.4	0.8
Age	25.4	3.5	24.4	1.8
Observations	28		26	

To check how homogeneous the two groups are, four main variables are considered. In Table 6, the mean and standard deviation of the age of both groups reveals that the average age is higher by one year and higher variability is apparent for the No-BTS condition group. Table 6 shows that while the majority of participants of the No-BTS condition were from Europe, the Middle-East seems to be the dominating region for the BTS group. The gender distribution is fairly similar across both conditions, but the education level seems to differ as shown in Table 7. Most of the participants under the No-BTS condition had a master's degree, while the majority of the BTS completed a bachelor's program.

Table 6: Descriptive Statistics of Region

	No-BTS	BTS
Region	Frequency	Frequency
Middle-East	9	13
Europe	17	10
Asia	1	3
South America	1	0

Table 7: Descriptive Statistics Gender and Education Level

	No-BTS		BTS	
Variable	Frequency of A	Frequency of B	Frequency of A	Frequency of B
Gender (A=Female, B=Male)	12	16	12	14
Education Level (A=Bachelors, B= Masters)	13	15	17	9

4.2.3. Binomial Logistic Regression

To test whether or not there is a relationship between each question and SDS under both conditions six binomial logits were conducted. A binomial logistic regression was chosen in these cases due to the nature of the possible values of the dependent variables, which are limited to two. The dependent variable is changed each time for each question, while the independent variables condition, SDS, and the interaction of both were fixed across all six logits. The aim behind this analysis is to identify which of the questions has a significant effect on the final results. The significance between both conditions will be determined by calculating the interaction term of both independent variables. The logic behind it is that the interaction term expresses the difference in effect of SDS on SSQS between BTS and No-BTS conditions.

As shown in Table 8, the only variable that was significantly impacted by the claimed explanatory variables is racial involvement mild. Condition had a significant positive effect with a p-value of 0.045, while social desirability score exhibits a similar direction with a p-value of 0.039.

The interaction term shows a similar trend to the one in the ordered logistic regression, a negative impact with a p-value of 0.031. All other dependent variables did not exhibit any significant results at the 5% and 10% significance level. To make sure that the results were not completely driven by one variable, another regression analysis was conducted by performing a parallel analysis that excluded the racial involvement mild variable from the SSQS (Adjusted SSQS). The regression is available under Appendix B1.

Table 8: Binomial Logistic Regression

	Racial I. Mild		Racial I. Moderate		Online Dating		Empathy		Privacy Respect		Drug Use	
Condition	6.486*	(0.045)	-0.415	(0.918)	-1.389	(0.649)	-1.458	(0.614)	1.909	(0.449)	3.365	(0.324)
Social Desirability Score	0.316*	(0.039)	-0.038	(0.836)	0.035	(0.772)	-0.007	(0.945)	0.133	(0.196)	0.133	(0.213)
Condition X Social Desirability Score	-0.404*	(0.031)	-0.048	(0.829)	0.147	(0.377)	0.089	(0.583)	-0.185	(0.195)	-0.158	(0.399)
Constant	-4.2019	(0.091)	3.215	(0.326)	-2.126	(0.323)	1.425	(0.451)	-1.273	(0.459)	-0.865	
Observations	54		54		54		54		54		54	

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

4.2.4. Correlation

An initial analysis was conducted to test the correlation between SDS and the SSQS of each participant under both conditions (BTS vs No-BTS). A correlation coefficient is typically used to examine the relationship between two ordinal or interval variables. In order to use the Pearson's correlation coefficient, two conditions must hold.

1. Both variables need to be interval.
2. Both variables need to be normally distributed.

Since both SDS and SSQS are interval variables, all that needs to be done is to test for normality. A Shapiro-Wilk test can be used in this case. The H0 is that the distribution of the given data is equivalent to a normal distribution, to be rejected if contrary ($p < 0.05$). Table 9 shows the results of the Shapiro-Wilk tests for SDS and SSQS variables. Under both conditions the null hypothesis cannot be rejected for both variables, indicating that the data is normally distributed.

Table 9: Shapiro-Wilk test of normality

Condition	Variable	Shapiro-Wilk		
		W	Z	Prob>z
No-BTS	SDS	.943	1.125	.130
	SSQS	.999	-2.356	.991
BTS	SDS	.984	-1.545	.939
	SSQS	.989	-5.996	1.00

Now that we know that the assumptions to run a Pearson's correlation coefficient test is satisfied, we can carry on with the analysis. Table 10 shows the correlation between SDS and

SSQS. Under the No-BTS condition, the correlation is 0.49, and flattens out to be 0.03 under the BTS condition.

Table 10: Pearson Correlation Coefficients.

Variable 1	**Variable 2**		**Condition**	**Correlations**	**Difference No BTS-BTS**
SDS	SSQS	No BTS	0.49	0.46	
		BTS	0.03		
		BTS	-0.20		

Figure 2 demonstrates that the correlation has decreased from 0.49 to 0.03, and the slope dropped from 0.103 to 0.005. Since It is expected to find a correlation between the prevalence of socially desirable responses and the social desirability score, SDS is expected to explain the variability of the results. The calculation of the R-squared for the No-BTS condition can reveal how much variability is due to social desirability bias, subtracting the BTS condition R-squared after would explain the extent of which BTS diminished the effect of SDB. As shown in Figure 2, 24.2% of the variability in the No-BTS condition is explained by social desirability score compared to 0.1% in the BTS condition. The difference of 24.1% indicates that BTS method almost entirely eradicated the effect of SDB.

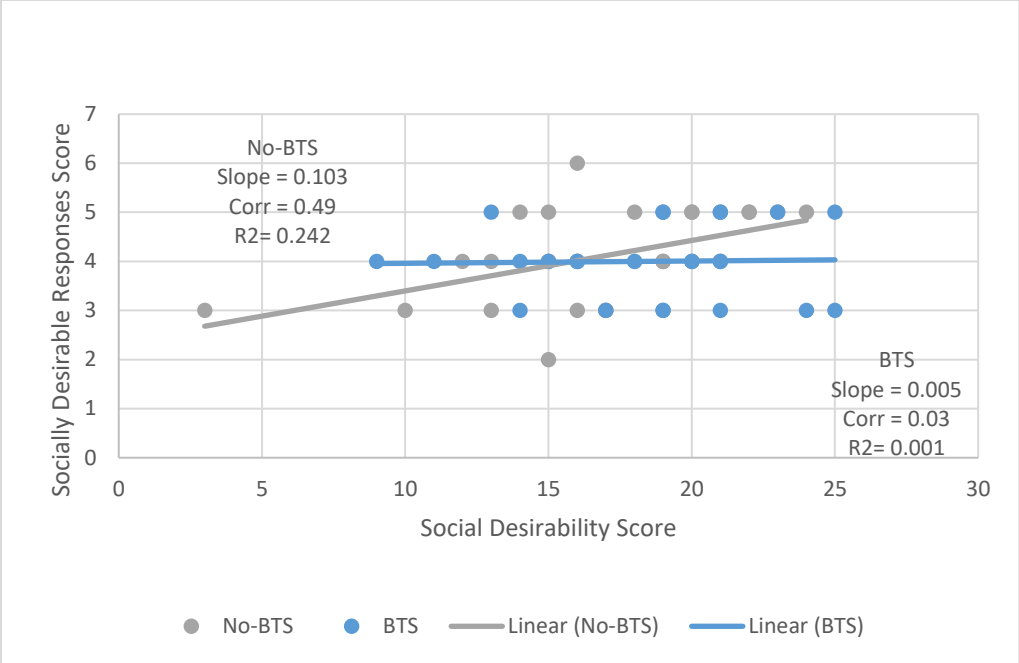


Figure 2: Correlation and slope for BTS and No-BTS conditions.

Since the correlation is calculated based on a sample extracted from the population, it is possible that we can obtain a correlation value different than zero by chance even if there is no linear relationship between the variables. Thus, it is important to know whether this is the case or not. The following hypotheses can be formed accordingly:

$$H_{0:}\, r = 0$$

$$H_1 : r \neq 0$$

According to Russo (2003), the appropriate t-test to use is:

$$t = \frac{r - 0}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

Where r is the correlation and n is the number of observations, and n-1 is the degrees of freedom. For the No-BTS condition we get the following result

$$t = \frac{0.49 - 0}{\sqrt{\frac{1 - 0.49^2}{28 - 2}}}$$

$$t = 2.866$$

The critical value in a t table for a p = 0.05(two tails) and 26 degrees of freedom is 2.06. The null hypothesis is to be rejected if the t score obtained is higher than the critical values for the given degrees of freedom. In this case we can reject the null hypothesis and declare that the sample was drawn from a population where SDS and SSQS are linearly associated variables.

For the BTS condition the t test result is as follows

$$t = \frac{0.03 - 0}{\sqrt{\frac{1 - 0.03^2}{26 - 2}}}$$

$$t = 0.147$$

The resulting t value is lower than the critical value with 24 degrees of freedom and p = 0.05, which is 2.06. Thus, we do not reject the null hypothesis.

4.2.5. Fisher R to Z Transformation test

To test whether the difference in correlation is significant or not, z test statistic can be conducted. Accordingly, we can form the following hypotheses for a *p*-value of 0.05.

$$H_0: r_1 = r_2$$
$$H_1: r_1 \neq r_2$$

Where r_1 is the correlation coefficient between SDS and SSQS for the No-BTS condition, and r_2 is the correlation coefficient between SDS and SSQS for the BTS condition.

First, the correlation coefficient values should be transformed into Z scores, which can be done through a Fisher's r to z transformation. The following equation can be used to conduct the transformation.

$$z = \frac{1}{2}\ln\left(\frac{1 + r}{1 - r}\right)$$

To get the z score for the No-BTS condition we compute the following

$$z_1 = \frac{1}{2}\ln\left(\frac{1+r_1}{1-r_1}\right) = \frac{1}{2}\ln\left(\frac{1+0.49}{1-0.49}\right) = 0.54$$

To get the z score for the BTS condition we compute the following

$$z_2 = \frac{1}{2}\ln\left(\frac{1+r_2}{1-r_2}\right) = \frac{1}{2}\ln\left(\frac{1+0.03}{1-0.03}\right) = 0.03$$

The next step is to compare z_1 and z_2, which can be done by computing the observed z and then get the corresponding *p*-value. The equation for the observed z is as follows:

$$z_{observed} = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

Where z_1 and z_2 are the z score for $r_1$and r_2, and $n_1$and n_2 are the number of observations for $r_1$and r_2 respectively. After plugging in all the values we get

$$z_{observed} = \frac{0.54 - 0.03}{\sqrt{\frac{1}{28 - 3} + \frac{1}{26 - 3}}} = 1.76$$

The corresponding *p*-value for $z_{observed}$ is 0.078. Accordingly, we cannot reject the null hypothesis that both correlations are equal at a 5% significance level. However, we can conclude that they are different at a 10% significance level.

4.2.6. Ordered Logistic Regression

An ordered logistic regression is conducted with SSQS as the dependent variable, while SDS and Condition are independent variables. Those independent variables were chosen to confirm the accuracy of the Marlow-Crowne social desirability scale and see whether or not the condition has a statistically significant effect on the final outcome. An ordered logistic regression was used since the dependent variable can have more than two values presented in the scale from 0 to 6, and the values have an ordered meaning embedded into them. The regression analysis is mainly conducted to investigate the effect of BTS on SDB, not in terms of absolute values of predictions but the change in trend between the BTS and No-BTS conditions. The proportional

odds assumption did not hold in this case since the BTS method disrupts the relationship between SDS and SSQS as hypothesized. When the ordered regression is conducted with SDS as the only explanatory variable, the model has a higher explanatory power and the parallel regression assumption holds. Since we are interested in how BTS decreases the correlation between SSQS and SDS, we will not be going with the model with the highest log-likelihood ratio but with the one that explores this effect. As a result, not meeting the proportional odds assumption is an expected byproduct and should not be overwhelming since we are not interested in accurate predictions but the change of predictions between both conditions.

All 54 observations in the dataset were used in the analysis. The likelihood ratio chi-square of 9.76 with a p-value of 0.0207 reveals that the model as a whole is statistically significant at a 5% significance level, compared to a model with no predictors.

As demonstrated in Table 11, social desirability score has a positive effect on SSQS indicating that an additional SDS point increases the probability to have an SSQS score of 6, ceteris paribus. The result is significant at the 1% level (p=0.003).

Table 11: Ordered Logistic Regression

SSQS	Coefficients	*p*-value
Condition	3.972	(0.079)
Social Desirability Score	0.262**	(0.003)
Condition X Social Desirability Score	-0.250*	(0.048)
Observations	54	

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

As for the interaction term between condition and SDS, a significant negative effect is observed at the 5%-level with a p-value of 0.048. The SDS in isolation demonstrates a positive effect on SSQS, which is moderated by the inclusion of the condition variable; altering the effect towards a negative direction.

For proper interpretation, from the interaction term, we can conclude that being in the BTS condition decreases the likelihood to end up with an SSQS score of 6 as SDS increases relative to No-BTS condition, with a 5% significance level, ceteris paribus. Appendix B3 and B4 presents the margins and the marginal effects of the current regression analysis.

4. Discussion

5.1. General discussion

In the light of concerns regarding the validity of self-report measures, this study was designed to provide a solution to overcome one of the major challenges (the effect of social desirability bias). The correlational study indicates that the application of BTS method almost completely eliminated the distortion that may have been caused by social desirability bias. When BTS method is applied, this relationship seems to fade out dropping from 0.49 to 0.03 as shown in Table 3 and Figure 1. A possible conclusion to derive out of this effect is that Bayesian Truth Serum Method diminishes the effect of social desirability bias from a moderate positive correlation to no correlation, and this effect is significant at a 10% level as the Fisher r to Z transformation analysis indicates. The ordered logistic regression emphasizes the statistical significance of this effect and provides a trend comparison between the degree of accuracy of SDS as an explanatory variable of SSQS and how SDS fails to explain variation when BTS is applied. The social desirability score in isolation demonstrates a positive effect on the social sensitivity questionnaire score, which is moderated by the inclusion of the condition variable; altering the effect towards a negative direction.

Since this is one of the first studies to investigate the use of BTS method to decrease the effect of social desirability bias, the six questions used were meant to cover a broad range of socially sensitive domains in an attempt to provide an exploratory input to be used for future research. In addition, the degree of social sensitivity was an additional concern. This was taken into considerations when deciding which questions to use in the social sensitivity questionnaire, more regarding this topic will be covered in the limitations section. Even though the results overall were significant, it was driven by one variable (Racial Involvement Mild). As a robustness check, another regression that excluded this variable from the SSQS was conducted and available under Appendix B1. It revealed that the results become insignificance once the variable was excluded. This might indicate two different things: either this specific variable was influenced by an unaccounted for effect or the other variables were.

Two main points of concerns might be raised. The first is whether the question used is actually considered a socially sensitive one. The second point is whether the answer claimed as

the socially desirable one, is in fact more socially desirable than the other. This reason is more complicated due to the heterogeneity of the group of participants, in terms of nationality, cultural background, gender, and education level, causing conflicting views regarding which response is more socially desirable.

To determine which of the claimed causes might have affected each of the variables, a determinative criterion can be used for each.

Claim #1: The question is socially sensitive to begin with.

To check if the question is socially sensitive or not, since SDS has proven to be a relatively stable variable to predict socially desirable responses, test for significance in correlation between the MCSDS score, and the frequency of socially desirable responses. This results with two criterion points on the checklist. In the case of any of the variables does not satisfy both conditions, the related question could be considered not socially sensitive.

1. Significant positive correlation between SDS and SDR (socially desirable response).
2. SDR frequency of No-BTS > SDR frequency of BTS.

Empathy and Drug use related questions are the suspects in this case. Both questions exhibit an abnormal trend to what previous studies has shown. There is no significant positive correlation between each of the variables and SDS. The application of BTS should increase truthfulness, decreasing the frequency of socially desirable answers. In this study, the frequency of socially desirable responses for empathy and drug use increased from 78.6% to 80.8% and 88.5% respectively. It was expected that people would be more ashamed to say that they care more about abused animals than starving children since evolutionary behavioral theory dictates that it is more natural to feel greater empathy towards who shares a greater genetic makeup with ourselves. On a second thought, it might be the opposite case since increased level of education lead to a less egoistic realization, that all life is precious and equal. Regarding drug use, even though it was expected to be a frowned down upon kind of behavior, it might be the case that for a one-time user it is a justifiable act of curiosity.

The online dating related question fails to meet both criteria but reveals an interesting observation. Since the expected trend was strictly reversed having 17.9% of socially desirable

responses in the No-BTS condition and 46.2% for the BTS group. This effect entertains the possibility that the question is in fact a socially sensitive question, but the chosen response is not the socially desirable one.

Claim #2: The chosen socially desirable response (SDR) is in fact the actual SDR.

Since the prevalence of socially desirable answers represent those who actually believe in their responses in addition to those who lie about it, and people have the tendency to respond in a socially desirable manner not the contrary. I will be using a rule of thumb here to consider the more frequent response as the socially desirable one. Accordingly, the first criterion is set. Regarding the second, the same logic from the first claim will be used. If the first condition is satisfied, we can conclude that the chosen SDR is in fact the SDR for the majority of the participants. If the second condition is satisfied, it would be fair to conclude that the chosen SDR for the entire group is the actual SDR. If not, then the further investigations to be carried out under the third claim should reveal if any of the demographic variables might explain why the chosen SDR is not universal across all groups.

1. SDR is chosen by the majority.
2. Significant positive correlation between SDS and SDR.

Reconsidering the alternative response as the socially desirable response for Online Dating, 82% responded in a socially desirable manner and this amount decreased to 53.8% under the BTS condition. However, the significance does not change since the binary outcomes were simply switched. To test for the effect of the modified online dating variable on the overall model, further regressions were made. Appendix B.3 shows additional ordered logistic regression that had the modified version of the online dating variable. Instead of considering "no" as the socially desirable response, these models included the alternative. As a result, the overall significance of the model increased for the SDS, condition, and interaction variables.

To summarize, the only variable that met both conditions is the Racial Involvement Mild. Excluding empathy and drug use due to not meeting the criteria of the first claim, there are only four variables left to question. Racial Involvement Mild met all criteria for claim two but racial involvement moderate, privacy respect, and online dating failed to meet the second criteria having no significant relationship between SDR and SDS. For Claim #3, only those four variables will be considered to see if the demographic variables might have led to the significance for Racial Involvement Mild or no significance for the others.

Claim #3: The frequency of SDR differ between regions, gender, and education level.

This claim is aimed to investigate whether there is a difference between demographic characteristics regarding which answer is perceived as socially desirable. It is tested for:

1. A significant difference between Middle-East and Europe.
2. A significant difference between Bachelors and Masters holders.
3. A significant difference between Females and Males.

Now since the criteria are set to test for the aforementioned claims that may have affected the results, let us go through each of the variables and check if any of the variables are affected by any of the claims. In this study, the tests for significance will not be conducted due to the limited number of observations under each of the demographic characteristics, however, the difference in frequency will be considered to draw some conclusions.

Regarding Racial Involvement Mild, the frequency of SDR is 86.4% for participants from the Middle East and 51.9% for Europeans. This might indicate that social desirability bias is more prominent in the Middle East or that the other response is the actual SDR in their culture. Since the difference in average SDS is only less than one point lower for the Middle-East group (18.27) relative to Europe (17.29), and the standard deviation is similar (4.2), the first conclusion will not be taken into further considerations. Since the amount of SDR is much lower among the Europe group, it might be an indicator that the other response is the SDR or the question is less socially sensitive in Europe. The same trend was found when comparing Bachelors to Masters graduates, but since most Europeans in the sample have a masters and most people from the Middle East have

bachelors, a case of multicollinearity was suspected. As a result, I will only be considering one of both, region variable. There was no significant difference between males and females in terms of SDR frequency.

Regarding racial involvement moderate a similar analysis was conducted to try and identify what other possible exogenous variables that may have affected the results. Even though the racial involvement moderate follows the expected trend, the majority choosing the socially desirable answers and the frequency percentage decreases when BTS is applied, the frequency of answers between people from the Middle East and Europeans show an interesting trend. All 27 Europeans answered in a socially desirable manner, while only 14 out of 22 people from the Middle East responded accordingly. Observing frequency difference of close to 36.36% raises the question whether the chosen socially desirable response is in fact the same for Europeans and Middle-Easterns. There was also a difference in frequency when considering gender indicating that it might be less socially acceptable for women to have a child with a person from a different ethnicity than men; since the frequency of SDR for females is significantly higher than males. Another conclusion would be that females tend to fall into social desirability bias more than males. The difference of the average SDS was 18.75 for females and 16.63 for males.

Privacy Respect seems to have a similar result when comparing frequency of responses between people from the Middle East and Europeans. For people from the Middle East, 40.9% chose the socially desirable answer and 59.1% did not. As for Europeans, 66.6% chose the socially desirable response and 33.3% did not. This almost reversed proportions can be explained by the conclusion that both groups do not perceive the socially acceptable answer similarly.

For online dating, 54.4% of respondents from the Middle East regions provided the socially desirable response while only 14.81% from the Europe region did so. A number of possible conclusions can be derived out of these frequencies. starting with the Middle East, since the prevalence of SDR is close to 50%, this might indicate that the questions is not as socially sensitive as thought. For European respondents, it might be the case that it is more socially desirable to not be against online dating rather than against. When considering both regions, the difference of 39.59% might indicate that Europeans are more open to online dating relative to people from the Middle East.

The influence of social desirability bias on self-reports depends on four main dimensions that formulate the intensity of the motive to provide a socially desirable response that deviates from reality. The four dimension are the distance between the socially desirable and undesirable response, the degree of sensitivity of the questions, the dispositional characteristics of the respondent, and the situational characteristics the respondent is facing (Donaldson and Grant-Vallone, 2002).

The distance between the socially desirable behavior and the not socially desirable one would be for example if the participant was asked to report his recycling habit on a 5 point scale, (1) never to (5) always. Always is the optimal, and the more the participant deviates from this response the more likely the respondent would be biased in his response. However, in this study this dimension was eliminated since all questions had two choices, either a socially desirable response or not. The degree of sensitivity is basically how socially sensitive the question is, as sensitivity level increases the likelihood to respond untruthfully in a socially desirable manner increases. This study included a combination of mildly to moderately socially sensitive questions; however, the degree of sensitivity was not of a particular interest in this study. The situational characteristics can be anything related to the external environment of the respondent. For example, place of answering the question (lab or home), was the participant being observed while answering, what is the consequences of not choosing the socially desirable response (will the participant be judged by the experimenter, or get fired for instance.

The dispositional characteristics refer to the characteristic of the respondent him or herself. More specifically in relation to SDB, the respondents propensity to provide a socially desirable answer. This is the point of interest in this study, and this propensity to provide SDR is measured by the MCSDS. Given that distance between the desirable and undesirable response did not play a role in this study, all respondents were exposed to the same level of sensitivity of questions, assuming that all of them respondent from the comfort of their home with identical circumstances, we can claim that BTS decreases SDB by tackling the dispositional characteristics dimension. We cannot know for sure if the difference in the number of SDR between one participant and the other is explicitly due to dispositional characteristics, since the situational characteristics and perception of degree of sensitivity was not controlled for.

Due to this study, we know that BTS acts in a similar manner to indirect questioning in regards to decreasing SDB and that incentives were added to induce truth telling. Interesting recommendations for future research would be comparing BTS and indirect questioning methods and try and determine which technique is more effective to reduce SDB. BTS can have two conditions, one with incentives and one without which may help determine and isolate the role of incentives and the method separately.

5.2. Limitations

Even though this study has led to some interesting insights, it is important to reflect on the limitations and shortcomings of the current research. To begin with, the tradeoff between incentive compatibility and anonymity was present in the current study. As seen earlier, to ensure anonymity the amount each participant gained was donated on their behalf to a charity for their choice. Even though this ensures anonymity, the donation part can lead to self-selection since it will attract people who are more lenient towards altruistic behavior than the average of the population. This might be problematic in this study, if this behavior is correlated with more socially desirable behavior than average. In this case, the deviation between the average amount of socially desirable responses and the actual responses from the samples in this study might not be strictly due to deception but due to the way participants in this sample generally behave. In addition, the donated amount might not be convincing enough to induce truth-telling. The maximum amount gained per truthful answer was around 25 cents, limited funding simply did not allow affording higher incentives. So direct incentives or a higher amount could have decreased the effect of self-selection.

Another major limitation that possibly influenced the results drastically, is the subjectivity of experimental criteria in the domain of socially desirable responses. Determining which answer is socially desirable was relatively challenging in this case. Initially, it was determined based on intuition before gathering the data or running the analysis. In the current study, the intuitive choices were still used, but was concluded to be questionable after noticing how evenly heterogeneous the observations were. This is considered to be a major limitation to the study. Possible solutions would be:

1. having a homogeneous group in terms of nationality, gender, and age which in turn can lead to external validity concerns as the sample is not representative to the population.
2. sophisticated criteria to determine if and how socially sensitive the questions are. This can be thought of as a more objective measure to determine which answer is to be considered the socially desirable one. The criteria used under the three claims discussed under the general discussion section can be considered an example.
3. an additional questionnaire asking participants which answer they consider more socially desirable themselves and to match these answers with the personal responses that were provided before. This can be done on an individual level, or a group level in terms of nationality as an example.

Lastly, sample size was not large enough according to the power calculations conducted under Appendix A.2. The achieved power with the given number of observations per group, 28 observations for No-BTS and 26 for BTS, is 0.42. To achieve a power of 0.8, 65 observations are required per group. This amount was hard to reach due to time and funding limitations. Thus, it might be the case that a large enough sample size was not obtained to make support the conclusions regarding the obtained results with absolute certainty.

5. Conclusion

The aim of the current study was to test whether the application of BTS method decreases the prevalence of social desirability bias in self-report questionnaires. The results revealed that BTS reduces the effect of SDB on self-report questionnaires. Noticing a diminishing correlation between the Marlowe-Crowne Social Desirability Score and the Social Sensitivity Questionnaire indicates the possibility that BTS can qualify as a suitable method to counteract the effect of social desirability bias. Even though the results were significant at a minimally accepted level (10%), it indicates that an additional study that accounts to the mentioned limitations should be able to provide conclusive evidence whether that is the case or not.

Most of the knowledge related to human behavior comes from subjective data gathered through self-report measures. Any study that includes socially sensitive questions is a potential victim of social desirability bias. The commonality of the prevalence of social desirability bias can lead to reporting false and misleading findings. Prior studies have indicated that social desirability bias can diminish, moderate, or inflate relationships between variables (Zerbe and Paulhus, 1987); influence variable means (Peterson and Kerin, 1981); and increase error in measurements (Cote and Buckley, 1988). Accounting for SDB and relying on proven methods to counter its effect is thus crucial to evade these mistakes. The results of this study indicate that BTS method is a potentially suitable candidate to rely on and get more valid and reliable subjective data.

References

Adams, S., Matthews, C., Ebbeling, C., Moore, C., Cunningham, J., Fulton J. and Herbert J. 2005. The effect of social desirability and social approval on self-reports of physical activity. American Journal of Epidemiology, 161(4):389-398.

Angner, E. (2012). *A course in behavioral economics*. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.

Ariely, D. (2008). *Predictably irrational: The hidden forces that shape our decisions*. New York, NY: Harper.

Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. Journal of Personality and Social Psychology, 46(3), 610-620. doi:10.1037//0022-3514.46.3.610

Bjerring, J. C., Hansen, J. U., & Pedersen, N. J. (2014). On the rationality of pluralistic ignorance. *Synthese, 191*(11), 2445-2470. doi:10.1007/s11229-014-0434-1

Brase, G. L. (n.d.). How different types of participant payments alter task performance. Udgment and Decision Making, 4(5), 419-428.

Crandall, R. (1976). Validation of Self-Report Measures Using Ratings By Others. *Sociological Methods & Research, 4*(3), 380-400. doi:10.1177/004912417600400305

Cote, J. A., & Buckley, M. R. (1988). Measurement Error and Theory Testing in Consumer Research: An Illustration of the Importance of Construct Validation. Journal of Consumer Research, 14(4), 579. doi:10.1086/209137

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*(4), 349-354. doi:10.1037/h0047358

Dawes, R.M., 1990. The potential non-falsity of the false consensus effect. In: R.M. Hogarth (Ed.), Insights in Decision Making. A Tribute to Hillel J. Einhorn University of Chicago Press, 179–199.

Del Missier, F., Ferrante, D., & Costantini, E. (2007). Focusing effects in predecisional information acquisition. Acta Psychologica, 125, 155-174.

Edwards, L. Allen, (1957). Techniques of Attitude Scale Construction. Vakils Feffer and Simons, Bombay.

Edwards, A. L., Diers, C. J., & Walker, J. N. (1962). Response sets and factor loadings on sixty-one personality scales. Journal of Applied Psychology, 46(3), 220-225. doi:10.1037/h0040280

Fisher, R. J. (1993). Social Desirability Bias and the Validity of Indirect Questioning. *J CONSUM RES Journal of Consumer Research, 20*(2), 303. doi:10.1086/209351

Halbesleben, J. R., & Bowler, W. M. (2007). Emotional exhaustion and job performance: The mediating role of motivation. Journal of Applied Psychology, 92(1), 93-106. doi:10.1037/0021-9010.92.1.93

Heider, F. The psychology of interpersonal relations. New York: Wiley, 1958.

Holmes, D. S. Dimensions of projection. Psychological Bulletin, 1968, 69, 248-268.

Huang, C., Liao, H. and Chang, S. 1998. Social desirability and the Clinical Self-Report Inventory: methodological reconsideration. Journal of Clinical Psychology, 54(4):517-528.

Kamenica, E. (2012). Behavioral Economics and Psychology of Incentives. *Annual Review of Economics Annu. Rev. Econ., 4*(1), 427-452. doi:10.1146/annurev-economics-080511-110909

Katz, D., & Allport, F. Students' attitudes. Syracuse: Craftsman Press, 1931.

Killian, L. M., Festinger, L., Riecken, H. W., & Schachter, S. (1957). When Prophecy Fails. American Sociological Review, 22(2), 236. doi:10.2307/2088869

King, M. and Bruner, G. 2000. Social desirability bias: a neglected aspect of validity testing. Psychology and Marketing, 17(2):79–103.

Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. Journal of Personality and Social Psychology, 67(4), 596-610. doi:10.1037//0022-3514.67.4.596

Lambert, C. E., Arbuckle, S. A., & Holden, R. R. (2016). The Marlowe–Crowne Social Desirability Scale outperforms the BIDR Impression Management Scale for identifying fakers. Journal of Research in Personality, 61, 80-86. doi:10.1016/j.jrp.2016.02.004

Murray, H. A. The effect of fear upon estimates of the maliciousness of other personalities. Journal of Social Psychology, 1933. 4, 310-339.

Nederhof, A. 1985. Methods of coping with social desirability bias: a review. European Journal of Social Psychology, 15(3):263-280.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175-220. doi:10.1037//1089-2680.2.2.175

Olshausen, B. A. (2004, March 1). Bayesian Probability Theory.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. Journal of Personality and Social Psychology, 46, 598-609.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), Measures of personality and social psychological attitudes (pp. 17-59). New York: Academic Press.

Paulhus, D. L. (1998). Paulhus Deception Scales (PDS) user's manual. North Tonawanda, NY: Multi-Health Systems.

Pelham, B., Sumarta, T., & Myaskovsky, L. (1994). The Easy Path From Many To Much: the Numerosity Heuristic. Cognitive Psychology, 26(2), 103-133. doi:10.1006/cogp.1994.1004

Peterson RA, Kerin RA. The quality of self-report data: review and syn-thesis. In: Enis BM, Roering KJ, editors. Review of marketing. Chicago(IL): American Marketing Association; 1981. p. 5 – 20.

Prelec, D. (2004). A Bayesian Truth Serum for Subjective Data. *Science, 306*(5695), 462-466. doi:10.1126/science.1102081

Prentice, D. A., & Miller, D. T. (1993). Pluralistic ignorance and alcohol use on campus: Some consequences of misperceiving the social norm. Journal of Personality and Social Psychology, 64(2), 243-256. doi:10.1037/0022-3514.64.2.243

Raghubir, P., & Menon, G. (1996). Asking sensitive questions: The effects of type of referent and frequency wording in counterbiasing methods. Psychology and Marketing, 13(7), 633-652. doi:10.1002/(sici)1520-6793(199610)13:7<633::aid-mar1>3.0.co;2-i

Raghubir, P., & Srivastava, J. (2002). Effect of Face Value on Product Valuation in Foreign Currencies. Journal of Consumer Research, 29(3), 335-347. doi:10.1086/344430

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. doi:10.1016/0022-1031(77)90049-x

Russo, R. (2003). Statistics for the behavioural sciences: an introduction. London: Psychology Press.

Shafir, E., Diamond, P., & Tversky, A. (1997). Money Illusion. The Quarterly Journal of Economics, 112(2), 341-374. doi:10.1162/003355397555208

Weaver, R., & Prelec, D. (2013). Creating Truth-Telling Incentives with the Bayesian Truth Serum. *Journal of Marketing Research, 50*(3), 289-302. doi:10.1509/jmr.09.0039

Zerbe, W. J., & Paulhus, D. L. (1987). Socially Desirable Responding in Organizational Behavior: A Reconception. Academy of Management Review, 12(2), 250-264. doi:10.5465/amr.1987.4307820

Appendix

A. Experiment

A.1. Questionnaire as presented to respondents

MCSDS

Section 1:

This section contains 33 statements concerning personal attitudes and traits. Read each item and decide whether the statement is *true* or *false* as it pertains to you personally.

Before voting I thoroughly investigate the qualifications of all the candidates.

True

False

I never hesitate to go out of my way to help someone in trouble.

True

False

It is sometimes hard for me to go on with my work if I am not encouraged.

True

False

I have never intensely disliked anyone.

True

False

On occasions I have had doubts about my ability to succeed in life.

True

False

I sometimes feel resentful when I don't get my way.

True

False

I am always careful about my manner of dress.

True

False

My table manners at home are as good as when I eat out in a restaurant.

True

False

If I could get into a movie without paying and be sure I was not seen, I would probably do it.

True

False

On a few occasions, I have given up something because I thought too little of my ability.

True

False

I like to gossip at times.

True

False

There have been times when I felt like rebelling against people in authority even though I knew they were right.

True

False

No matter who I'm talking to, I'm always a good listener.

True

False

I can remember "playing sick" to get out of something.

True

False

There have been occasions when I have taken advantage of someone.

True

False

I'm always willing to admit it when I make a mistake.

True

False

I always try to practice what I preach.

True

False

I don't find it particularly difficult to get along with loudmouthed, obnoxious people.

True

False

I sometimes try to get even rather than forgive and forget.

True

False

When I don't know something I don't mind at all admitting it.

True

False

I am always courteous, even to people who are disagreeable.

True

False

At times I have really insisted on having things my own way.

True

False

There have been occasions when I felt like smashing things.

True

False

I would never think of letting someone else be punished for my wrong-doings.

True

False

I never resent being asked to return a favor.

True

False

I have never been irked when people expressed ideas very different from my own.

True

False

I never make a long trip without checking the safety of my car.

True

False

There have been times when I was quite jealous of the good fortune of others.

True

False

I have almost never felt the urge to tell someone off.

True

False

I am sometimes irritated by people who ask favors of me.

True

False

I have never felt that I was punished without cause.

True

False

I sometimes think when people have a misfortune they only got what they deserved.

True

False

I have never deliberately said something that hurt someone's feelings.

True

False

>>

SSQS BTS

Section 2:

Answer the following 6 two-part questions as indicated.

For each question you answer you will receive an "information score". The sum of all information scores will then be calculated to come with your overall "Truth Score."

Truth scoring, recently invented by an MIT professor and published in the academic journal Science, rewards you for answering truthfully. Even though only you know if you really answered truthfully or not, people who tell the truth score higher overall.

You are most likely to maximize your potential donations if you answer every item truthfully. By "truthfully," we mean: consider each question carefully, answer honestly, and take care to avoid mistakes.

As mentioned earlier, truthful answers are the best way to maximize the "Truth Score." This will then determine the amount of money to be donated on your behalf to an Egyptian charity organization or project of your choice. Just by participating, EGP 5 (Egyptian Pounds) will be donated. This amount can increase to up to EGP 20 based on your "Truth Score." In other words, each truthful answer will contribute with a maximum of EGP 2.50.

Based on your "Truth Score", you will gain a certain amount of money that will then be transferred to a charity organization under the project that you choose. The Available projects are as follows, please select your preferred project:

[Kiosk Project (sustainable income)] It aims to provide a poor family with a source of a steady income, in this case the source is a Kiosk fully equipped, licensed, and contains initial goods valued at EGP 3000. The establishment of each kiosk costs the organization EGP 8000 and is estimated to provide a monthly income of EGP 500.

[Winter Blankets Project] Aims to help the most vulnerable families in villages and communities to face the winter cold by providing blankets during the coming winter season. Each blanket costs around 30 EGP.

[The Children's Cancer Hospital Foundation (CCHF) 57357] The CCHF 57357 is a legal independent non-profit organization with a vision "to be the unique worldwide icon of change towards a cancer-free childhood."

If you were going to have a child, would it be a problem if the other parent wouldn't be of the same ethnicity as you?

Yes

No

Out of 100 respondents, how many do you think will answer "Yes" and how many will answer "No" to the same question?

Make sure Yes + No = 100

0 10 20 30 40 50 60 70 80 90 100

Yes

No

Would you be willing to meet someone through an online dating app or site?

Yes

No

Out of 100 respondents, how many do you think will answer "No" and how many will answer "Yes" to the same question?

Make sure Yes + No = 100

0 10 20 30 40 50 60 70 80 90 100

No

Yes

According to you, which is worse: starving children or abused animals?

Starving children

Abused animals

Out of 100 respondents, how many do you think will answer "starving children" and how many will answer "abused animals" to the same question?

Make sure Starving Children + Abused Animals = 100

0 10 20 30 40 50 60 70 80 90 100

Starving children

Abused animals

Would you—for any reason—read your mate's email without his/her knowledge and permission?

Yes

No

Out of 100 respondents, how many do you think will answer "No" and how many will answer "Yes" to the same question?

Make sure Yes + No = 100

0 10 20 30 40 50 60 70 80 90 100

No

Yes

Have you used psychedelic drugs (LSD, mescaline, peyote, etc.) ?

Yes

No

Out of 100 respondents, how many do you think will answer "Yes" and how many will answer "No" to the same question?

Make sure Yes + No = 100

0 10 20 30 40 50 60 70 80 90 100

Yes

No

Would you prefer to go out with someone of your own skin color / racial background?

Yes

No

Out of 100 respondents, how many do you think will answer "No" and how many will answer "Yes" to the same question?

Make sure Yes + No = 100

0 10 20 30 40 50 60 70 80 90 100

No

Yes

Please indicate your age in the space below.

Please indicate your gender.

Male

Female

SSQS No-BTS

Section 2:

Please answer the following 6 questions as indicated.

By completing this section, EGP 20 (Egyptian Pounds) will be donated to an Egyptian charity Project or Organization of your choice. Please indicate your choice below:

[Kiosk Project (sustainable income)] It aims to provide a poor family with a source of a steady income, in this case the source is a Kiosk fully equipped, licensed, and contains initial goods valued at EGP 3000. The establishment of each kiosk costs the organization EGP 8000 and is estimated to provide a monthly income of EGP 500.

[Winter Blankets Project] Aims to help the most vulnerable families in villages and communities to face the winter cold by providing blankets during the coming winter season. Each blanket costs around 30 EGP.

[The Children's Cancer Hospital Foundation (CCHF) 57357] The CCHF 57357 is a legal independent non-profit organization with a vision "to be the unique worldwide icon of change towards a cancer-free childhood."

If you were going to have a child, would it be a problem if the other parent wouldn't be of the same ethnicity as you?

Yes

No

Would you be willing to meet someone through an online dating app or site?

Yes

No

Which is worse: starving children or abused animals?

Starving children

Abused animals

Would you—for any reason—read your mate's email without his/her knowledge and permission?

Yes

No

Have you used psychedelic drugs (LSD, mescaline, peyote, etc.) ?

Yes

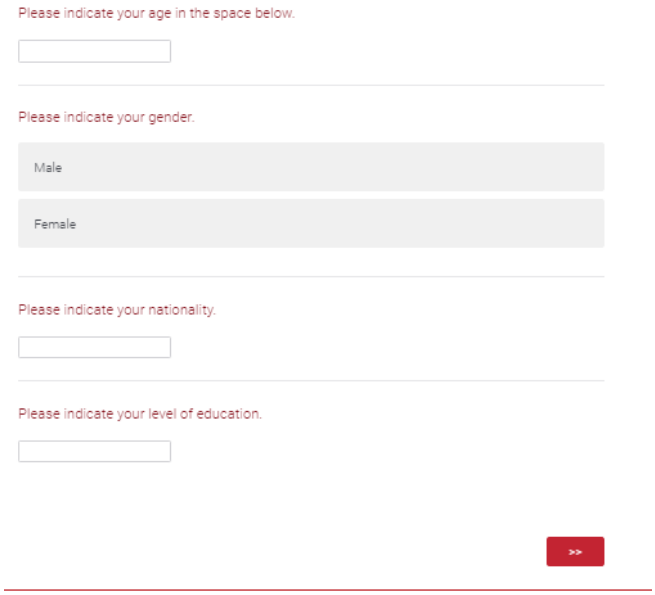
No

Would you prefer to go out with someone of your own skin color / racial background?

Yes

No

Demographic Questions



A.2. Power Calculations

To test whether the sample sized used was optimal or not, a two independent samples power correlations calculations can be conducted. The commonly used significance value of alpha = 0.05 will be used. Given correlation coefficients of $r_1 = 0.49$ and $r_2 = 0.03$, and sample sizes of $n_1 = 28$ and $n_2 = 26$, we have a power value of $1 - \beta = 0.42$ as shown in Figure (3). In order to reach a power level of 0.8, we need to have 65 observations for each conditions as shown in Figure (4), which was not attainable in this study due to time and funding limitations.

```
. power twocorrelations 0.49 0.03, alpha(0.05) n1(28) n2(26)

Estimated power for a two-sample correlations test
Fisher's z test
Ho: r2 = r1  versus  Ha: r2 != r1

Study parameters:

        alpha =    0.0500
            N =        54
           N1 =        28
           N2 =        26
        N2/N1 =    0.9286
        delta =   -0.4600
           r1 =    0.4900
           r2 =    0.0300

Estimated power:

        power =    0.4175
```

Figure 3 : Screen shot of Stata output to calculate acquired power.

```
. power twocorrelations 0.49 0.03, alpha(0.05) beta(0.2)

Performing iteration ...

Estimated sample sizes for a two-sample correlations test
Fisher's z test
Ho: r2 = r1  versus  Ha: r2 != r1

Study parameters:

        alpha =    0.0500
         beta =    0.2000
        delta =   -0.4600
           r1 =    0.4900
           r2 =    0.0300

Estimated sample sizes:

            N =       130
  N per group =        65
```

Figure 3 : Screen shot of Stata output to calculate required sample size to achieve desired power level..

Given the sample size, the power of the correlation comparison can be calculated. The needed input is the correlations of both groups, available observations, and the alpha in this case it is 0.05.

A.3. BTS and Bayesian Nash Equilibrium

This part of the section is aimed to clarify how it is in the respondent's best interest to answer truthfully. Let us begin by brief introduction to Nash equilibrium in relation to Bayesian theory. Nash equilibrium is a stable state that involves the interaction of various agents, in which no agent achieves the optimal gain through a unilateral strategic change given a constant unchanged strategic behavior from the remaining agents. A Bayesian Nash Equilibrium situation is formed in the case of BTS since the respondent has the option of telling the truth or deceiving. Given that all other respondents will be telling the truth as instructed, respondent r is better off behaving accordingly. Let us consider the same question from, as presented below. Following the example from "Prior and Bayesian Updating" subsection before, if respondent r is choosing answer $x^r \epsilon \{0,1\}$ (Y=1), and determine frequency $y^r \epsilon \{0,1\}$ of yes respondents.

In the case of answering yes ($x^r = 1$ and y^r), the respondent score looks as follows:

$$\left[log\frac{\bar{x}}{\bar{y}}\right] + \left[\bar{x} log\frac{y^r}{\bar{x}} + (1-\bar{x}) log\frac{(1-y^r)}{(1-\bar{x})}\right]$$

In the case of answering yes ($x^r = 1$ and y^r), the respondent score looks as follows:

$$\left[log\frac{(1-\bar{x})}{\underline{y}}\right] + \left[\bar{x} log\frac{y^r}{\bar{x}} + (1-\bar{x}) log\frac{(1-y^r)}{(1-\bar{x})}\right]$$

$$\bar{x} = average\ of\ all\ x^r$$

$$\bar{y} = geometric\ average\ of\ all\ y^r$$

$$\underline{y} = geometric\ average\ of\ all\ (1-y^r)$$

The first term is the information score and it is not dependent on y^r.As for the second term, it is the prediction score and is not dependent on x^r.

Truth-telling in the case of BTS is Bayesian Nash Equilibrium, since if all respondents are telling the truth the optimal outcome for respondent r is to behave accordingly. Respondents r expects 75% of other respondents to answering yes ($x^r = 1$), and 25% of repondents to answer no ($x^r = 0$).

$\bar{x} = E(\omega | t_r = Y) = 0.75$.

Respondents r would then think that 75% of respondents in Y category will report a prediction of $y^r = 0.75$, and the other 25% of respondents will report $y^r = 0.67$. As a result

$$\bar{y} = 0.75^{0.75} * 0.67^{0.25} = 0.73$$

$$\underline{y} = 0.25^{0.75} * 0.33^{0.25} = 0.27$$

Since we have all the values to compute the information score for a Y respondent.

If answered yes ($x^r = 1$):

$$iScore = log\frac{\bar{x}}{\bar{y}} = log\frac{0.75}{0.73} = 0.012$$

If answered yes ($x^r = 0$):

$$iScore = log\frac{(1-\bar{x})}{\underline{y}} = log\frac{0.25}{0.27} = -0.033$$

The results show that a respondent that belongs to category Y is better off answering yes ($x^r = 1$) as it maximizes the information score.

Even though the prediction score was not used in this study by giving alpha a value of 0, I will carry on with the calculation to provide an example of how it would work. In order to determine which prediction maximizes the prediction score, we simply need to compute the first derivative of the prediction score, with respect to y^r.

$$\frac{0.75}{y^r} - \frac{(1-0.75)}{(1-y^r)} = 0$$

$$0.75 - 0.75y^r = y^r - 0.75y^r$$

$$0.75 = y^r = \bar{x}$$

The results imply that the prediction score is maximized when the provided prediction (y^r) exactly matches the average of all responses ($\bar{x}$).

B. Data, Analysis and Results

B.1. Adjusted Ordered Logistic Regression

Table 12 shows the results of the ordered logit using the adjusted as the dependent variable. After removing the Racial Involvement Mild from the SSQS, the effect condition and the interaction does not exhibit any significance regarding these variables. On the other hand, the SDS has a positive effect on the adjusted SSQS, increasing the probability to get a score of 5, with a significance at 10% (p-value = 0.072), ceteris paribus.

Table 12: Adjusted Ordered Logistic Regression

Adjusted SSQS	Coefficients	*p*-value
Condition	0.775	(0.738)
Social Desirability Score	0.145*	(0.072)
Condition X Social Desirability Score	-0.058	(0.656)
Observations	54	

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

B.2. Modified Online Ordered Logistic Regression

Table 13 shows the results of the ordered logit using the Online modified SSQS dependent variable. After considering yes as the socially desirable response to the online dating question, the effect condition, the SDS, and the interaction does exhibit significance. SDS has a positive effect on the SSQS, increasing the probability to get a score of 6, with a significance at 5% (p-value = 0.017), ceteris paribus. The condition has a significant positive effect at a 5% level aswell with a p-value of 0.04. The interaction term shows a significant negative effect at a 1% level (p-value = 0.006).

Table 13: Ordered Logistic Regression (Online Variable Modified)

Socially Desirable Reponses Score	Coefficients	*p*-value
Condition	4.387*	(0.040)
Social Desirability Score	0.194*	(0.017)
Condition X Social Desirability Score	-0.33**	(0.006)
Observations	54	

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

B.3. Marginal effect

Prior to calculating the marginal effect of SDS, and the interaction, the only predicted outcome that showed significance at a 5%-level for SDS and interaction is an SSQS score of 5. Using these variables, we can try and figure out if an additional SDS point increases the probability to get an SSQS score of 5 by the same amount for BTS and No-BTS conditions.

Looking at Table 14, we can conclude that on average, an additional SDS point increases the probability to have an SSQS outcome of 5 by 4.6 percentage points, ceteris paribus, at a 1% significance level. As for the interaction term, on average, an additional SDS point decreases the probability to have an SSQS score of 5 by 4.4 percentage points, at a 5% significance level, ceteris paribus. This indicates that an additional SDS point increases the probability to have an outcome of 5 by 4.4 percentage points less under the BTS condition relative to the No-BTS condition. Taking this into account, we can conclude that on average an additional SDS point increases the probability to have an SSQS outcome of 5 by 4.6 percentage points without treatment (No-BTS) and 0.2 percentage points with treatment (BTS). Based on these results, we can confirm that an additional SDS point does not increase the probability to have an SSQS outcome of 5 by the same amount for the BTS and No-BTS conditions.

Table 14: Average Marginal Effects

	SSQS 2		SSQS 3		SSQS 4		SSQS 5		SSQS 6	
Condition	-0.201	(0.454)	-0.255	(0.234)	0.033	(0.193)	0.191	(0.348)	0.233	(0.349)
Social Desirability Score	-0.004	(0.277)	-0.041**	(0.004)	-0.006	(0.588)	0.046**	(0.001)	0.004	(0.323)
Condition X Social Desirability Score	0.004	(0.313)	0.039	(0.055)	0.005	(0.587)	-0.044*	(0.037)	-0.004	(0.353)
Observations	54		54		54		54		54	

p-values in parentheses

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

B.4. Margins

Earlier it was concluded that on average, the probability of having an outcome of 5 increases by 4.6 percentage points with an additional SDS point under the No-BTS condition and 0.2 under the BTS condition.

Table 15: Margins

		SSQS = 3		SSQS = 4		SSQS = 5	
SDS	**Condition**	**Margin**	***p*-value**	**Margin**	***p*-value**	**Margin**	***p*-value**
14	**0**	0.351	0.001	0.453	0.000	0.168	0.017
14	**1**	0.257	0.009	0.476	0.000	0.243	0.012
15	**0**	0.297	0.002	0.471	0.000	0.207	0.005
15	**1**	0.255	0.004	0.476	0.000	0.245	0.005
16	**0**	0.247	0.002	0.476	0.000	0.252	0.001
16	**1**	0.252	0.002	0.476	0.000	0.247	0.002
17	**0**	0.203	0.003	0.470	0.000	0.303	0.000
17	**1**	0.250	0.001	0.476	0.000	0.249	0.001
18	**0**	0.164	0.007	0.451	0.000	0.358	0.000
18	**1**	0.248	0.001	0.476	0.000	0.251	0.001
19	**0**	0.132	0.016	0.422	0.000	0.415	0.000
19	**1**	0.246	0.002	0.476	0.000	0.253	0.001
20	**0**	0.105	0.032	0.385	0.000	0.473	0.000
20	**1**	0.244	0.004	0.476	0.000	0.255	0.003
21	**0**	0.083	0.059	0.343	0.000	0.529	0.000
21	**1**	0.242	0.008	0.476	0.000	0.258	0.006
22	**0**	0.065	0.095	0.299	0.003	0.580	0.000
22	**1**	0.240	0.018	0.476	0.000	0.260	0.014

This effect can be observed in Table 15 under the "SSQS = 5" column. It indicates that the probability to end up with a SSQS outcome of 5 is 0.303 when having an SDS score equal to 17 and no BTS condition (0), the probability increases to 0.358 when SDS is equal to 18. The difference in probability is 5.5%, around the same average predicted increase. As for the increase under the BTS condition (1), when SDS is 17 the predicted probability is 0.249 and increases to 0.251, a marginal effect of 0.002 is observed. All the mentioned values are significant at a 1% level. The difference in the marginal increase indicates that when BTS is applied, the effect of SDS on SSQS decrease and remain constant. This effect is clearly indicated in Figure (3)

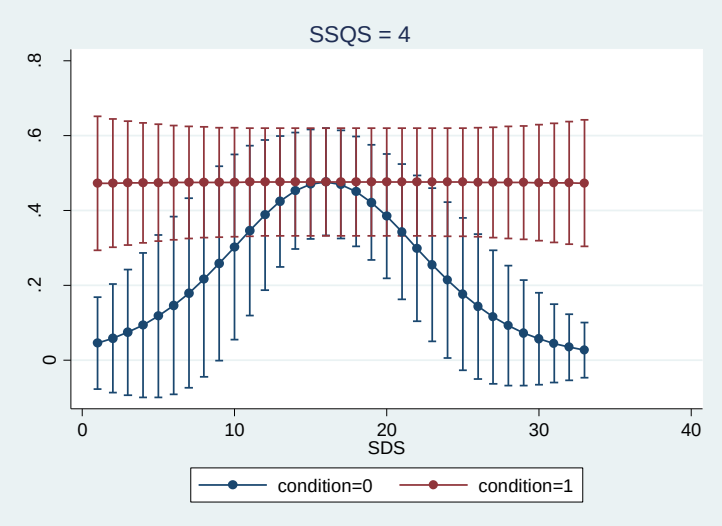


Figure 3: Margins Graph of Outcome 4

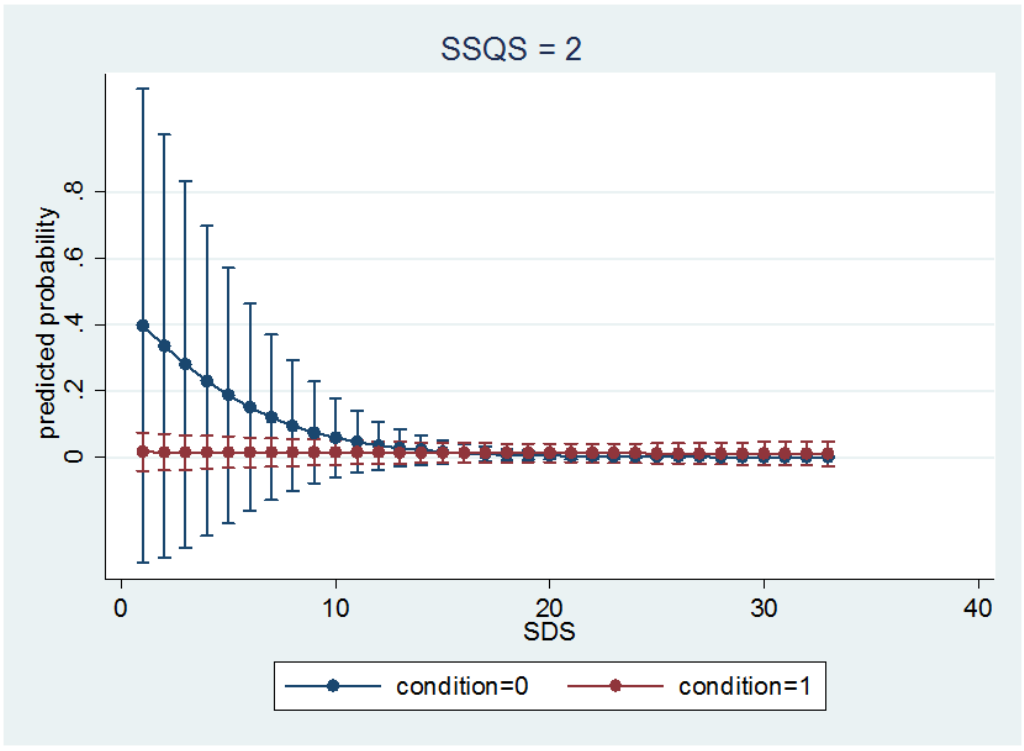
The average SDS and SSQS is 17.6 and 4.1, respectively. This is indicated in Figure 3 with the highest predictability of an outcome of 4 is when SDS is around 18. Under no BTS condition, probability diminishes quite drastically as the SDS deviates from this point. As for the BTS, it is interesting to notice how the predicted probability of outcome 4 remains fairly constant across the SDS scale, indicating that the BTS method eliminated the effect of social desirability bias. This trend is observable across all possible SSQS outcomes as presented in Appendix B.6, along with the full table of margins under Appendix B.5.

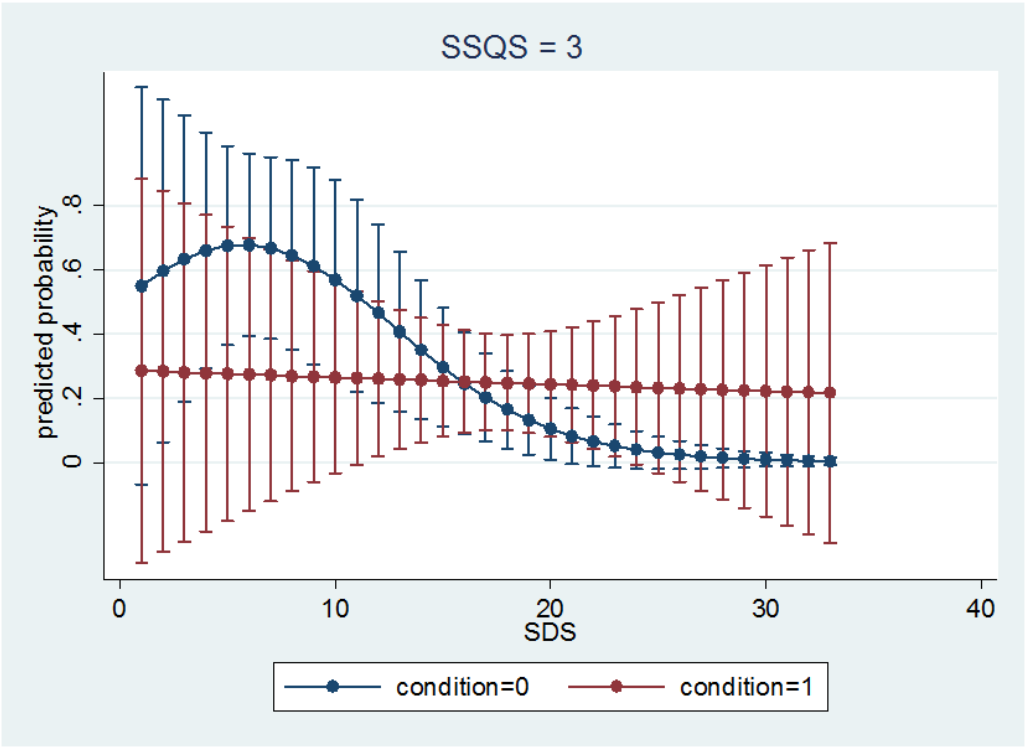
B.5. Ordered Logistic Regression Full Margins Table

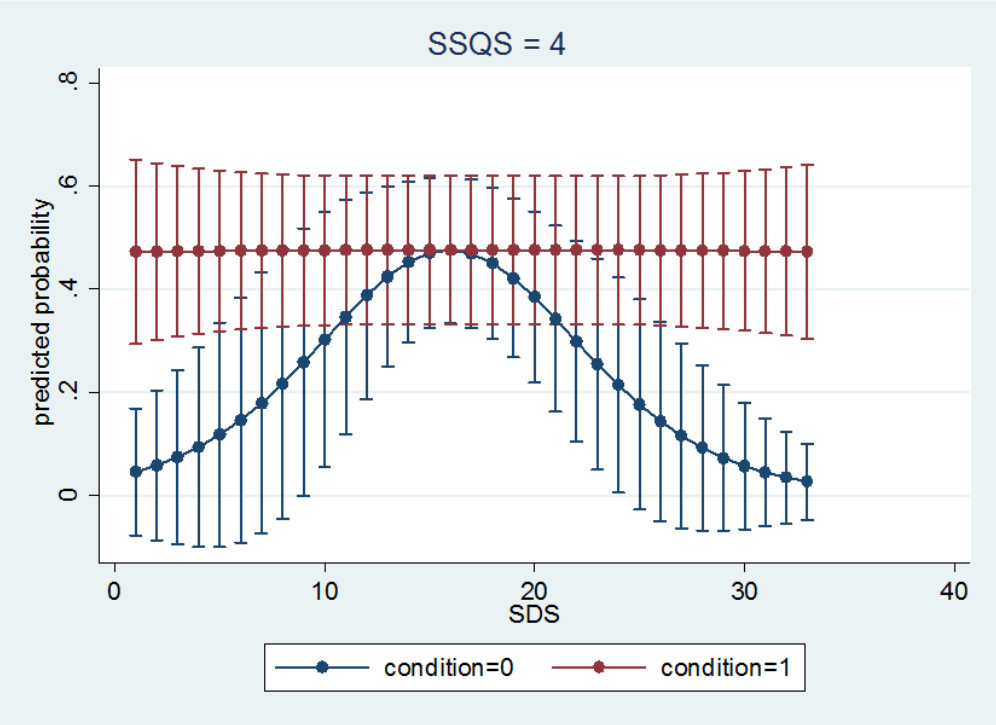
		SSQS = 2		SSQS = 3		SSQS = 4		SSQS = 5		SSQS = 6	
SDS	**Condition**	**Margin**	***p*-value**	**Margin**	***p*-value**	**Margin**	***p*-value**	**Margin**	***p*-value**	**Margin**	***p*-value**
1	**0**	0.397	0.276	0.550	0.081	0.046	0.462	0.007	0.510	0.000	0.600
1	**1**	0.016	0.594	0.286	0.348	0.473	0.000	0.216	0.404	0.009	0.590
2	**0**	0.336	0.302	0.596	0.028	0.059	0.426	0.009	0.484	0.000	0.585
2	**1**	0.015	0.580	0.284	0.322	0.473	0.000	0.218	0.376	0.010	0.575
3	**0**	0.280	0.321	0.633	0.005	0.075	0.385	0.011	0.456	0.000	0.569
3	**1**	0.015	0.566	0.282	0.295	0.474	0.000	0.220	0.346	0.010	0.559
4	**0**	0.231	0.334	0.660	0.000	0.094	0.339	0.015	0.425	0.001	0.553
4	**1**	0.015	0.550	0.279	0.266	0.474	0.000	0.222	0.313	0.010	0.543
5	**0**	0.187	0.341	0.675	0.000	0.118	0.286	0.019	0.390	0.001	0.536
5	**1**	0.015	0.535	0.277	0.236	0.474	0.000	0.224	0.280	0.010	0.526
6	**0**	0.151	0.344	0.678	0.000	0.146	0.227	0.025	0.353	0.001	0.518
6	**1**	0.015	0.519	0.275	0.205	0.475	0.000	0.226	0.244	0.010	0.509
7	**0**	0.120	0.344	0.668	0.000	0.180	0.165	0.032	0.311	0.001	0.500
7	**1**	0.015	0.502	0.272	0.173	0.475	0.000	0.228	0.208	0.010	0.491
8	**0**	0.095	0.342	0.646	0.000	0.217	0.103	0.041	0.266	0.001	0.481
8	**1**	0.014	0.486	0.270	0.141	0.475	0.000	0.230	0.170	0.010	0.473
9	**0**	0.075	0.339	0.612	0.000	0.259	0.051	0.052	0.219	0.002	0.462
9	**1**	0.014	0.469	0.268	0.110	0.475	0.000	0.232	0.134	0.010	0.455
10	**0**	0.059	0.336	0.570	0.000	0.302	0.017	0.067	0.169	0.002	0.442
10	**1**	0.014	0.453	0.266	0.080	0.476	0.000	0.234	0.099	0.011	0.437
11	**0**	0.046	0.334	0.520	0.001	0.346	0.003	0.085	0.120	0.003	0.423
11	**1**	0.014	0.438	0.263	0.055	0.476	0.000	0.236	0.068	0.011	0.420
12	**0**	0.036	0.334	0.465	0.001	0.388	0.000	0.108	0.076	0.004	0.404
12	**1**	0.014	0.424	0.261	0.034	0.476	0.000	0.238	0.043	0.011	0.403
13	**0**	0.028	0.336	0.408	0.001	0.425	0.000	0.135	0.040	0.005	0.386
13	**1**	0.014	0.411	0.259	0.019	0.476	0.000	0.240	0.024	0.011	0.389
14	**0**	0.021	0.340	0.351	0.001	0.453	0.000	0.168	0.017	0.007	0.370
14	**1**	0.013	0.399	0.257	0.009	0.476	0.000	0.243	0.012	0.011	0.376
15	**0**	0.017	0.347	0.297	0.002	0.471	0.000	0.207	0.005	0.009	0.355
15	**1**	0.013	0.390	0.255	0.004	0.476	0.000	0.245	0.005	0.011	0.365
16	**0**	0.013	0.357	0.247	0.002	0.476	0.000	0.252	0.001	0.012	0.342
16	**1**	0.013	0.384	0.252	0.002	0.476	0.000	0.247	0.002	0.011	0.357
17	**0**	0.010	0.368	0.203	0.003	0.470	0.000	0.303	0.000	0.015	0.332
17	**1**	0.013	0.380	0.250	0.001	0.476	0.000	0.249	0.001	0.011	0.352
18	**0**	0.008	0.382	0.164	0.007	0.451	0.000	0.358	0.000	0.019	0.325
18	**1**	0.013	0.379	0.248	0.001	0.476	0.000	0.251	0.001	0.012	0.351
19	**0**	0.006	0.398	0.132	0.016	0.422	0.000	0.415	0.000	0.025	0.321
19	**1**	0.013	0.381	0.246	0.002	0.476	0.000	0.253	0.001	0.012	0.352

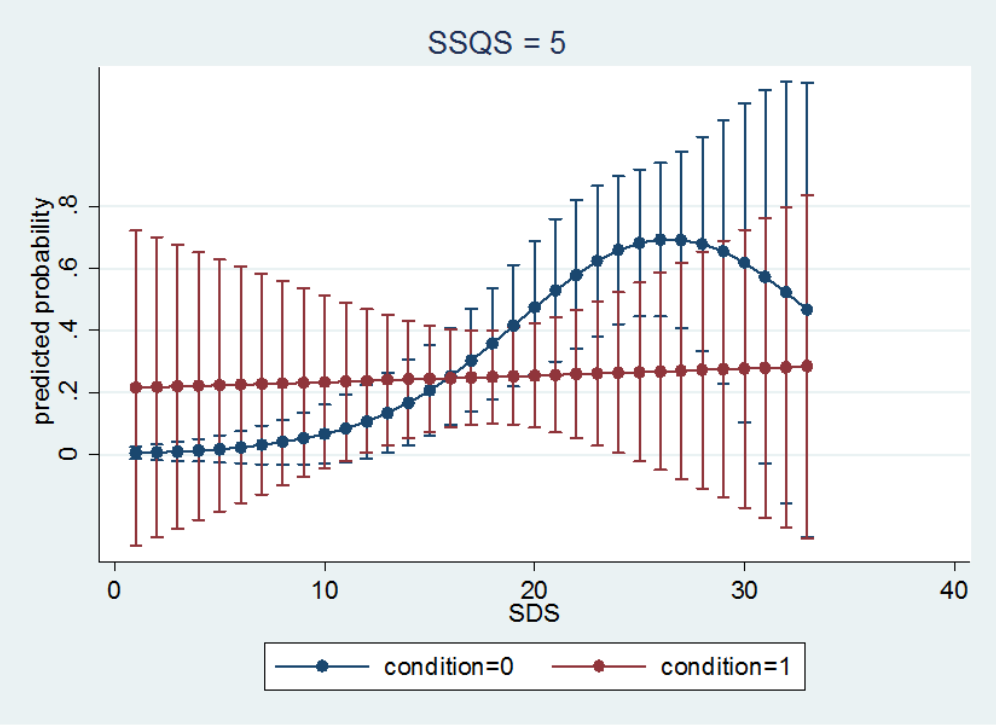
20	**0**	0.005	0.415	0.105	0.032	0.385	0.000	0.473	0.000	0.032	0.319
20	**1**	0.013	0.385	0.244	0.004	0.476	0.000	0.255	0.003	0.012	0.357
21	**0**	0.003	0.433	0.083	0.059	0.343	0.000	0.529	0.000	0.042	0.320
21	**1**	0.012	0.393	0.242	0.008	0.476	0.000	0.258	0.006	0.012	0.365
22	**0**	0.003	0.451	0.065	0.095	0.299	0.003	0.580	0.000	0.054	0.323
22	**1**	0.012	0.402	0.240	0.018	0.476	0.000	0.260	0.014	0.012	0.376
23	**0**	0.002	0.469	0.051	0.138	0.255	0.015	0.623	0.000	0.069	0.327
23	**1**	0.012	0.414	0.238	0.034	0.476	0.000	0.262	0.026	0.012	0.389
24	**0**	0.002	0.488	0.040	0.184	0.214	0.044	0.657	0.000	0.087	0.331
24	**1**	0.012	0.427	0.235	0.056	0.476	0.000	0.264	0.045	0.012	0.403
25	**0**	0.001	0.506	0.031	0.230	0.177	0.089	0.681	0.000	0.111	0.335
25	**1**	0.012	0.442	0.233	0.085	0.476	0.000	0.267	0.069	0.013	0.420
26	**0**	0.001	0.524	0.024	0.275	0.144	0.145	0.692	0.000	0.139	0.337
26	**1**	0.012	0.458	0.231	0.118	0.475	0.000	0.269	0.097	0.013	0.437
27	**0**	0.001	0.541	0.019	0.318	0.116	0.203	0.691	0.000	0.173	0.337
27	**1**	0.012	0.474	0.229	0.154	0.475	0.000	0.271	0.127	0.013	0.455
28	**0**	0.001	0.557	0.014	0.357	0.092	0.259	0.678	0.000	0.214	0.332
28	**1**	0.011	0.491	0.227	0.191	0.475	0.000	0.273	0.159	0.013	0.473
29	**0**	0.000	0.573	0.011	0.393	0.073	0.310	0.654	0.003	0.262	0.323
29	**1**	0.011	0.507	0.225	0.228	0.475	0.000	0.276	0.191	0.013	0.490
30	**0**	0.000	0.588	0.009	0.426	0.058	0.357	0.618	0.018	0.315	0.307
30	**1**	0.011	0.523	0.223	0.264	0.474	0.000	0.278	0.223	0.013	0.508
31	**0**	0.000	0.602	0.007	0.457	0.045	0.398	0.574	0.061	0.374	0.284
31	**1**	0.011	0.540	0.221	0.298	0.474	0.000	0.280	0.254	0.013	0.526
32	**0**	0.000	0.616	0.005	0.484	0.035	0.435	0.522	0.131	0.437	0.254
32	**1**	0.011	0.555	0.219	0.331	0.474	0.000	0.283	0.283	0.014	0.542
33	**0**	0.000	0.629	0.004	0.510	0.027	0.468	0.466	0.212	0.503	0.216
33	**1**	0.011	0.570	0.217	0.362	0.473	0.000	0.285	0.311	0.014	0.559

B.6. Ordered Logistic Regression Margins Graphs






SSQS = 4
predicted probability
0
.2
.4
.6
.8
0
10
20
30
40
SDS
condition=0
condition=1


SSQS = 5
predicted probability
0
.2
.4
.6
.8
0
10
20
30
40
SDS
condition=0
condition=1


SSQS = 6
predicted probability
0
.2
.4
.6
.8
0
10
20
30
40
SDS
condition=0
condition=1