Erasmus School of Economics

Thesis

To obtain the academic degree of

Master of Science in Economics & Business

(Major in Behavioral Economics)

Exploring the discrepancy between willingness-to-pay and willingness-to-accept with the Bayesian Truth Serum

Author: Marina Georgieva

Email address: 414358mg@eur.nl

Supervisor: Han Bleichrodt

Study Program: Business Economics

Specialization: Behavioral Economics

Date: 23/08/2016

Abstract

The willingness-to-pay and willingness-to-accept divergence has been explored for the last 40 years. Researchers have investigated both the explanations for the discrepancy as well as possible ways of reducing it. One research direction of decreasing the WTA/WTP ratio is by implementing incentive-compatible methods within the experimental design. This study tests a relatively new incentive-compatible method, the *Bayesian truth serum*, over the traditionally employed non-incentivized introspection in an attempt to eliminate the disparity between WTA and WTP for a new product provided by Uber. BTS rewards truth telling by assigning each answer a truth score, given respondent's endorsement and a prediction of the distribution of others' answers. For the chosen research context, an altered BTS version is proposed and tested. The results indicate that the method does eliminate the discrepancy, however, removing outliers chance the result's direction significantly suggesting that BTS is inferior. Discussed are future research directions for improvements of the BTS method and implications.

Table of Contents

List of figures

Introduction

Standard economic theory suggests that the difference between individuals' maximum willingness-to-pay (WTP) to obtain a good and the minimum willingness-to-accept (WTA) to trade a good should be insignificant (Willig, 1976). However, for the past 40 years scholars have reported on multiple occasions unexpected divergence between WTP and WTA (Brown, 2005). In their review of over 50 studies, Horowitz & McConnell (2002) report a WTA/WTP mean ratio of 2.9 for "ordinary private goods" and 10.4 for non-marketed public goods. Brown & Gregory (1999) observe a mean ratio of 2.5 for 13 studies involving non-environmental goods and 16.6 for 10 studies about environmental goods.

While a stream of research focuses on the explanations for the disparity (Kahneman, Knetsch, & Thaler, 1991; Horowitz & McConnell, 2002;Brown & Gregory, 1999; Zhao & Kling, 2001), others have examined how it can be minimized (Sayman & Öncüler, 2005; Coursey, Hovis, & Schulze, 1987; Shogren, Shin, Hayes, & Kliebenstein, 1994). Researchers argue that the discrepancy is due to poor experimental design and elicitation techniques (Brookshire & Coursey, 1987). It is suggested that individuals may intentionally give false valuation because of intrinsic motivations such as bargaining habits and capital gain motives (Knez, Smith, & Williams, 1985). Therefore, incentive-compatible methods should be employed to reduce the disparity by giving subjects reasons to reveal their true preferences. Popular methods used are the Vickrey's auction (Vickrey, 1961) and the BDM mechanism (Becker, DeGroot, & Marschak, 1964). However, empirical evidence regarding the effect of incentive-compatible methods on the WTA/WTP ratio is mixed. For instance, Horowitz & McConnell (2002) report that incentive-compatible designs increase the ratio, while in their meta-analysis from 39 studies, Sayman & Öncüler (2005) conclude that the ratio is decreased when subjects are incentivized to reveal their true valuations, signalling that the contrast might be caused by the mechanism itself.

The focus of this paper will be to investigate a relatively new incentive-compatible method: the Bayesian Truth Serum (hereafter referred to as BTS) (Prelec, 2004) and its effect on the WTA-WTP discrepancy. What makes the BTS different is its ability to elicit truthful responses even when the truth is unverifiable. It awards truthful responses for both rare and widely shared answers e.g. it does not depend on consensus as a norm for truth-telling. BTS's idea is to reward truthful answers by assigning them high scores. It uses both personal answers from respondents and their prediction of the distribution about the answers of others for estimating the "truth scores". A "surprisingly common" answers are rewarded as they receive high scores when the actual frequency is higher that its predicted frequency obtained from the same sample (Weaver & Prelec, 2012). For example, if an answer is endorsed by 10% of the respondents and the same

respondents predicted that its frequency will be 5%, then a high score is assigned to it (Prelec, 2004). On the other hand, "unsurprisingly common" answers are penalized.

BTS has not been widely utilized, however, its applications so far yield mostly positive results in terms of its validity (Weaver & Prelec, 2012; John, Loewenstein, & Prelec, 2012; Barrage & Lee, 2010; Loughran, Paternoster, & Thomas, 2014). Weaver & Prelec (2012) and Barrage & Lee (2010) investigate whether truth elicitation can be obtained in valuation of nonmarket goods. A domain in which the method has not been tested yet is "willingness-to-pay" and "willingness-to-accept" measurements for new marketed goods and services. Knowledge about WTA/WTP ratio is essential for marketed goods since its gap might cause error predictions for trade volume and actual gains resulting in 'market stickiness' (Borges & Knetsch, 1998).

An interesting domain for examining WTA/WTP would be the "shared economy" or collaborative consumption (CC hereafter). Investopedia defined CC as "an economic model in which individuals are able to borrow or rent assets owned by someone else"[1]. It provides an alternative of the traditional buyer-seller market by lending, swapping, sharing, bartering, trading and renting goods and services, usually provided by long-established companies (Botsman, 2010). Although not a new concept, CC has been growing rapidly in recent years and its fields are numerous: transportation, goods and services, accommodation, money, online content, etc. (Rosenberg, 2013). With the aid of the Internet, successful companies such as Airbnb, Uber, YouTube, Etsy, Kickstarter, e-Bay, etc. have flourished.

CC's main feature is the shift from ownership to temporary access (Bardhi & Eckhardt, 2012; Belk, 2013; Perren et al., 2015). This allows for underutilized assets to become means for increasing income (Cusumano, 2014). Additionally, consumers are no longer simply consumers; they become actively involved in the production process by being "co-creators" or "prosumers" (Denegri-Knott & Zwick, 2011).

Forbes estimated a revenue of $3.5 billion in 2013 for this sector which directly goes to the consumers (Geron, 2013), with a growth of 25% per year. CC platforms provide cost-effective goods and services and offer opportunities to its users to diversify their income by becoming micro-entrepreneurs (Schor, 2014). But what are the challenges? The 2015 1099 Economy Workforce Report revealed that working for on-demand companies as Uber results in high attrition rates as people's expectations about the payment and flexible working hours are not met. Some may argue that this new type of services is disruptive for established businesses and may not be sustainable (Schor, 2014; Leonard, 2014; Morozov 2014; Cusumano, 2014), thus, examining WTA/WTP ratio in this domain could be an important indicator of how the market

[1] Definition of 'Sharing Economy' by Investopedia

dynamics will unfold. Moreover, eliciting truthful answers can be an important predictor for actual gains and volume of services in this new domain.

The current research is focused on a new urban logistics product from Uber. Currently, Uber's main product is a taxi-like service in which drivers and riders are connected through a mobile application as a percentage of the fee goes to Uber and the rest for the drivers. The role of Uber is twofold: first, the application is a facilitator for drivers and rides; second, it is responsible for setting prices (Rogers, 2015). Thus, determining truthful WTP and WTA and reducing the gap between them is crucial for the company to set accurate and fair prices as to avoid high attrition rates among its drivers and customers. As the company's plans are to expand into the urban logistics services (Badger, 2014), truthfully defining WTA/WTP ratio could be used for proper forecasting of potential gains as well as market size for this domain.

The objective of this paper is to investigate whether responses acquired by BTS will minimize the ratio of WTA/WTP over non-incentivized traditional methods of assessing WTP and WTA, in an area where truthful answers do not exists at the time of the research, namely WTP and WTA for the new product provided by Uber in the Netherlands. The main motivation of this paper is to examine whether:

1) BTS can prove itself to be a more truthful elicitation technique in comparison with introspection;
2) BTS can be used as a measurement tool for WTP and WTA where differences among them are insignificant.

Currently, no research has focused on examining WTA/WTP ratio in the CC marketplace. One of the research's contributions will be in that area. More importantly, the study will investigate whether BTS can help overcome the disparity between WTP and WTA. Finally, the current work will contribute to BTS's application for marketed goods in assessing both WTP and WTA.

The thesis is structured in five chapters: 1) introduction; 2) literature review; 3) research methodology and data; 4) data analysis and results; and 5) conclusions and directions for future research. The literature review consists of an overview of the three major theoretical and research streams used to derive the main hypothesis, namely the WTA/WTP discrepancy, Bayesian Truth Serum and the sharing economy. In Chapter 3 presented are the altered version of the BTS's scoring mechanism and the data collection procedures. Chapter 4 consists of a detailed analysis of the main results and findings obtained from the performed investigations in order to answer the posed research questions. Chapter 5 presents the conclusions from the current study and discusses the main implications for both the academia and businesses. In addition, suggestions for future research and limitations are presented in this chapter.

Literature Review

The literature review will cover several research streams that are used for hypothesis derivation. First, discussed are the the WTP and WTA disparity, ways of reducing the ratio and the incentive-compatible methodologies applied in that domain. Second, described is the BTS, its intuition, approach and previous research. Finally, the focus is on the collaborative consumption as a context for testing the WTP and WTA with the BTS.

WTP and WTA disparity

While **standard economic theory** suggests that when there are small income effects, differences between individual maximum willingness-to-pay and minimum willingness-to-accept should be insignificant (Willig, 1976), scholars have found on numerous occasions a significant difference between one's selling price and one's buying price, the selling one being higher. WTP refers to the maximum amount that one is willing to pay to obtain a good, while WTA represents the minimum amount that one is willing to accept to sell the same good (Brown & Gregory, 1999). WTP is then a buying price, whereas WTA is a selling price.

Consumer theory proposes that the gap should not occur if: 1) there's no income effect, 2) no transaction costs; 3) perfect information about the product and its price and 4) means to truthfully reveal the valuation (Brown, 2005). However, with the required conditions in hand, scholars found that even for inexpensive goods the gap persisted. The impact of the discrepancy for consumer behaviour, thus will be reflected in two directions (Kahneman, Knetsch, & Thaler, 1990; Knetsch, 1989; Borges & Knetsch, 1998). First, the presence of the disparity decreases the volume of the trade that can be obtained than initially assumed. That is, potential traders might be unwilling to trade at prices higher (for buyers) or lower (for sellers) than expected. Consequently, the gains would be lower than estimated. Using two sets of costless market simulations, Borges & Knetsch (1998) showed that actual gains from the trade and volume are far less than conventionally assumed. Moreover, the authors estimated the number of Pareto-efficient[2] exchanges in a competitive market and concluded that there will be failure in allocating the goods to those willing to pay the most for them, thus welfare won't be optimal.

Researchers have been interested in both explanations and exploring ways to reduce the disparity. The reasons for the occurrence of the discrepancy have been widely studied. For example, Brown & Gregory (1999) classify them as **economical** (income effects and substitutes, transaction costs, implied value and profit motive) and **psychological** (endowment effect,

[2] When allocating resources, Pareto-efficiency occurs when it is impossible to make one party better off without making at least one person worse off.

legitimacy, ambiguity and responsibility). Kahneman, Knetsch, & Thaler (1991) use the endowment effect as an explanation for the diversion. The endowment effect refers to the "increased value of a good to an individual when the good becomes part of the individual's endowment" (Kahneman et al., 1990). It is corollary to loss aversion: losses loom larger than gains, therefore there's higher disutility of giving up a good than there is utility for obtaining the same good (Kahneman & Tversky, 1979). Simply put, people value losses more than they value gains. Furthermore, Zhao & Kling (2001) propose that "uncertainty, irreversibility and limited learning opportunities" evoke commitment costs which bring about the disparity. Hanemann (1991) suggests that the analytical approach towards determining an approximate equality of WTP and WTA has been "misconceived". His model takes into account the effect of available substitutes of the good on the WTA/WTP ratio. Holding the income effect fixed, larger gap would be expected when the good has fewer substitutes.

One direction of investigating the high WTA/WTP ratio is related to the **experimental design itself**. Horowitz & McConnell (2002) examine nearly 50 studies to seek reasons for the gap. Their main findings are: 1) hypothetical and non-incentive compatible do not yield higher ratios – that is, experiments using truth elicitation techniques actually lead to either no change or higher ratio; 2) lower ratios are present when the subjects are students; 3) as the good moves away from "ordinary private good", the ratio increases, e.g. ordinary goods have lower ratio than non-ordinary ones. Furthermore, Sayman & Öncüler (2005) investigate how the WTA/WTP ratio varies as a result of certain context related variables. In their meta-analysis, the examined variables are related to the payment methods, subject design and preference revealing mechanisms. They argue that these variables are perceived at different levels of importance for the buyer and the seller, thus their valuations might be affected. For example, one way to reveal the true valuations of subjects is by using *iterative bidding*. First, subjects are asked for their valuation. Afterwards, there are questions whether they will pay/accept a revised amount which is lower or higher than their initial answer. This iterative process allows for truthful revelation. Moreover, the researchers investigate how the subject design would impact the WTA/WTP ratio. Within-subjects design would result in lower ratio as subjects may try to be consistent with their answers. Adopting the idea of mental accounting (Thaler, 1985), the authors argue that as the different mental accounts for payments are based on where the money comes from and what is their purpose (i.e. where are they going), the different payment mechanisms can as well have an impact on participants' valuation. The authors found that out-of-pocket payments increase the disparity, while iterative bidding and within-subject designs decrease it.

Researchers also argue that the limited leaning opportunities embedded in surveys and experiments will increase the disparity and the **learning experience** should be utilized in the experimental design so that the divergence is reduced (Zhao & Kling, 2001; Coursey, Hovis, & Schulze, 1987; Shogren, Shin, Hayes, & Kliebenstein, 1994). Zhao & Kling (2001) argue that when

people make real-life decisions their final choice and valuation are based on information search and processing. The survey design doesn't provide such chances, rather subjects are forced to make a decision in an environment where they have not "voluntarily" stopped their information gathering resulting in gap rise. However, conclusions have been mixed in this domain. Using Vickrey's auction, Coursey et al. (1987) found that although the disparity for the first rounds of the auction is quite big, it almost disappears after series of trials. Thus, market-like experience is in line with standard economic theory. Shogren, Shin, Hayes, & Kliebenstein (1994) found that for marketed goods WTP and WTA do not differ after a certain number of trials, however for non-marketed goods the same results don't apply. On the other hand, Kahneman et al. (1990) didn't find significant decrease in the ratio when subjects repeated the task.

WTA/WTP ratio also differs for the **good** being studied. Hanemann (1991) showed that for products with fewer substitutes, the ratio is likely to be larger than for products with more substitutes. He claims that the difference in the measures depends on both the income and substitution effects and for products with almost no or imperfect substitutes the discrepancy will be higher. The intuition behind it is that the compensation for giving up a unique product can differ from the price for acquiring the same product when income is limited. Shogren et al. (1994) demonstrate that the divergence for public goods with no substitutes persists even when full information about the product is available. For marketed goods such as candy bars and coffee mugs, the discrepancy doesn't exist. This is consistent with Horowitz & McConnell (2002) studies' review which concludes that as the good moves away from being "private ordinary good", the gap increases. The difficulty when measuring non-marketed or public goods is to exert the "true" WTP and WTA. In such cases, incentive-compatible methods come in beneficial as the task requires truth elicitation. Thus, researchers have examined whether using **different incentive-compatible** methods reduces the ratio. Even though the rationale behind the usage of incentive-compatible methods is validated in theory, the literature shows no consensus whether their usage is justified in practice.

Experimental economists argue that when monetary incentives are present and dominate over other intrinsic or extrinsic factors, one has greater motives to reveal his or her true preferences as it will maximize the final payoff (Smith, 1982). When one is not incentivized to reveal his or her true preferences, the WTP and WTA will be exaggerated in 'opposite directions' (Brookshire & Coursey, 1987). In the WTP set-up the actual demand for a product will be understated, on the contrary, the valuation for compensation in the WTA set-up will be overstated as to reduce the product's supply. Therefore, without properly designing a mechanism for truthful elicitation the gap will persist.

Carson & Groves (2007) provide three reasons why survey-based questions yield results not in line with economic theory: 1) strategic misinterpretation; 2) hypothetical nature of the experiment and 3) not well defined preferences. Economic theory suggests that when surveyed, economic agents who believe their responses will have an impact on decisions taken by businesses or governments for outcomes that agents care about, should respond in a way as to maximize their expected welfare (Carson & Groves, 2007). However, as Scott (1965) noted "Ask a hypothetical question and you get a hypothetical answer." Hypothetical techniques provide little or no incentives for subjects to carefully think about their responses (Trautmann & van de Kuilen, 2014). Moreover, they can misinterpret their valuations due do different motives (e.g. social desirability bias) (Manski, 2004) which results in strategic misinterpretation. This strategic misinterpretation can cause potential problems in regards to the truthful revelation when no incentives are provided. Additionally, Loomis (2014) noted that what seems to be a hypothetical bias is actually problem concerting truthful elicitation. However, many studies suggest that hypothetical questions (introspection) work fine in many domains (L Guiso, Jappelli, & Terlizzese, 1992; Carman & Kooreman, 2011; Luigi Guiso & Parigi, 1999; Hurd, 2009). In terms of *accuracy*, the support for introspection against incentive-compatible mechanisms is mixed (Sonnemans, Kuilen, & Wakker, 2009; Rutström & Wilcox, 2009; Hollard, Massoni, & Vergnaud, 2010; Friedman & Massaro, 1998). For example, Trautmann & van de Kuilen (2014) have compared 4 "truth serums" each of different complexity with the standard non-incentivized introspection. Their results indicate that there is unnecessary complexity for some of the truth serums, making them inconvenient for researchers when there's time pressure, not enough funds or high number of questions asked. Thus, they find no benefit of using the truth serums against the introspection method.

Mixed conclusions have also been reported for the *validity* of incentive-compatible methods. As mentioned above, Horowitz & McConnell (2002) found that incentive-compatible methods such as the Vickrey auction, BDM mechanism and others actually yield higher WTA/WTP ratio. Counterintuitively, techniques designed to yield truthful answers provide either no change or an increase in the ratio. On the other hand, the notion that incentive-compatible methods result in lowering the ratio is supported in numerous researches (Coursey et al., 1987; Sayman & Öncüler, 2005; Brookshire & Coursey, 1987). For instance, Brookshire & Coursey (1987) measure the ratio WTA/WTP in a hypothetical environment and in an incentivized marketplace. What they found is that for hypothetical elicitation techniques the discrepancy exists, while in the market-like environment with incentives it decreases greatly. The authors argue that a market-like environment consists of two important characteristics that reduce the asymmetry between WTP and WTA: appropriate incentives for truthful revelation of demand and learning opportunities. Relevant incentives should be provided for truthful responses, once acquired they will account for any source of biases that can appear.

Several methods have been used for eliciting truthful (subjective) WTP and WTA valuations. For marketed goods the most popular ones include Vickrey Auction and Becker-deGroot-Marschak (BDM) mechanism. Vickrey's auction is a modification of the traditional sealed auction: the highest bidder wins and pays the second highest bid (Vickrey, 1961). However, the method is not created to induce "true" answers, but rather to show participants that this should be their dominant strategy (Coursey et al., 1987). The BDM method compares subjects' bids to a randomly generated price (Becker et al., 1964). If the bid is higher, the subject pays the randomly generated price and gets the item. If the bid is lower than the elicited price, the subject doesn't pay and receives nothing. Here again, the optimal strategy for one is to state their true preferences, otherwise there is risk for not acquiring the item being auctioned when the true preference doesn't coincide with the revealed.

There are some disadvantages concerning BDM mechanism and Vickrey's auction when exploring WTP and WTA for a new product. Most importantly, the methods are designed to resemble real-life purchase situation in which subjects are asked for the price they would pay/accept for acquiring/selling the tested product. However, people cannot be required to purchase a non-existent good and use the price as a benchmark for comparison (Loomis, 2014). Stating the true valuation would be the optimal strategy for the subjects, but the methods are not designed to measure the truth when is actually unknowable (e.g. WTP and WTA for new product at the time of the experiment). Moreover, researchers have questioned the incentive-compatibility of both methods as it has been shown that participants do not always follow the optimal strategy, rather they tend to understate or overstate their WTP and WTA (Karni & Safra, 1987; Kaas & Ruprecht, 2006; Horowitz, 2006; Berry, Fischer, & Guiteras, 2011). For example, Karni & Safra (1987) showed that BDM is not incentive-compatible when the value of the good in question is uncertain. Additionally, Horowitz (2006) demonstrated that the problem persists even if the value is certain. The main argument is that as the generated price is random, subjects are uncertain of the amount they will be required to pay, resulting in the possibility that one's bid is influenced by the distribution of prices. The outcome will then be not revealing the true valuation.

The drawbacks of these methods and the mixed evidence on whether the WTA/WTP ratio can be reduced by incentive-compatible methods, implies an underlying problem in the mechanisms. Given the importance of truthful valuations of WTA/WTP in consumer theory, in this paper I will investigate a new truth elicitation technique: Bayesian Truth Serum (Prelec, 2004). By comparing BTS with introspection, the aim of the research would be to verify whether indeed using incentive-compatible methods is justified. The paper will test if the BTS method can

help overcome the disparity between WTP and WTA and if it can be used to improve the quality of the data. In the next section, BTS intuition and approach are described.

Bayesian Truth Serum

Intuition

The applicability of the Bayesian Truth Serum in the WTA/WTP discrepancy relies on several characteristics of the model. First, the method is designed to measure the truth when it is in fact "unknowable" (Prelec, 2004). The challenge to acquire truthful responses when the truth is unobservable at the time of the study is essential when deriving WTA/WTP for new products. Second, it does not rely on consensus, it assures that unpopular valuations are not penalized and truth-telling is the best strategy. Third, subjects are not required to understand the scoring behind the method in order to maximize their scores. Unlike other mechanism, knowledge of BTS scoring rule is not beneficial for the respondents, e.g. they don't need to estimate their optimal strategy. Rather, subjects are being told that their best strategy should be to state their valuations thoughtfully and accurately.

Bayesian truth serum was proposed in 2004 by Drazen Prelec. BTS is a scoring method which provides incentives for truthful responses. Its basic idea is to appoint high score to an answer whose estimated frequency is higher than its predicted frequency collected from the same sample (Weaver & Prelec, 2012). Such answers are referred to as "surprisingly common". The "surprisingly uncommon" are the ones that have higher predicted frequency and are penalized by receiving low scores. Based on the scores the researcher can incentivize truth telling even when the truthful responses are unknown (Weaver & Prelec, 2012).

The intuition behind these high scores comes from a physiological phenomenon known as the "false consensus effect" (Ross, 1977): people rely on their own beliefs and perceptions about the world to draw conclusions about others' beliefs and perceptions. The authors conducted several experiments to find support for the false consensus bias in choices and behavior, personal traits such as habits, daily activities, expectations and opinions. For example, they found that students willing to wear sign in the university campus expected that 65% of other students would also wear it. However, students not willing to wear it expected only 31% of the other students to wear it. Moreover, optimistic students indicated that 62% of other students are also optimistic, while non-optimistic ones estimated the percentage to be 50. Generally, when one supports a statement he/she would expect the same behavior to be more common than for others who don't support the statement. Thus, the false consensus effect is an egocentric bias which suggests that people will overestimate the extent to which others' behave and think as they do. The name

of the effect suggests that this is irrational behavior: it is a bias which leads to false perception about a consensus that might not exist (is it possible that everyone is right?).

In 1989, Dawes challenged this notion of irrationality and proved that the effect is rational. He argues that people giving different answers have in fact different basis for estimating the proportions:

"When a prediction is to be made, the person who answers "yes" to a question has a basis for estimating a proportion that is different from that of the person who answers "no," and people who have access to different diagnostic data generally should-by normative reasoning-make different predictions about proportions and probabilities. Exactly how different is mathematically derivable."

He corrects for the failure of understanding that one's response serves as a "cue" for the responses of the population. Using Bayesian analysis, he shows that two opposite answers (e.g. yes and no) are "conditionally independent, and hence weighted equally". That is, when a group supports largely an opinion, behaviour, choice, etc. one is likely to fall into this group. When a small proportion of the group support an opinion, behaviour, choice one is not likely to fall into this group.

Prelec (2004) uses this notion as a critical assumption in his algorithm: people rely on their own beliefs to draw estimates about others' behaviour. Bayesian interpretation predicts that the estimates of the distribution of the frequency of an answer are inferred based on a prior personal belief. Bayesian reasoning tells us that one should expect that others will underestimate the true frequency of one's own opinion (Prelec, 2004). On average people underestimate the percentage of the frequency for opinions of others and overestimate the percentage of their own opinions (Weaver & Prelec, 2012). Therefore, one's own truth is more common than collectively predicted. This results in the fact that common answers will be the ones which have higher actual frequency than predicted by others which leads to the conclusion that truth can be obtained even when it's not verifiable.

Approach

BTS can be used in a multiple, binary or open choice questions where the respondents answer two questions (Prelec, 2004). First, the subject should give his or her personal opinion and then provide estimates of others' answers. Based on the inputs from both answers an information score is calculated. Its formula is as follows:

$$Information\ Score\ =\ log(\frac{actual\ frequency}{geometric\ mean\ of\ answer's\ predicted\ frequency})$$

For example, an answer will receive high score when the actual average frequency is higher than the predicted one. If one assumes that 5% of the population will support his or hers opinion and the actual frequency is higher than 5 (e.g. 10%), then high score is awarded for that answer as it is more common than collectively predicted. However, if the actual frequency is lower than 5% (e.g. 15%), then the answer gets lower score.

The total score of a respondent is a combination between the information score and the respondent's prediction score or the score for his accurate prediction:

$$Score\ for\ responded\ \boldsymbol{r} = \sum_k x_k^r\, log\, \frac{\bar{x}_k}{\bar{y}_k} + \alpha \sum_k \bar{x}_k log\, \frac{y_k^r}{\bar{x}_k}$$

It is symmetric and 0 if α=1. The first part is the information score and the second is the prediction score of respondent *r*. As noted above, the information score is high for respondents providing answers with actual frequency higher than predicted. The best prediction score is the one for which the actual frequency is the same as the predicted one ($\bar{x}_k$=y_k^r, e.g. it is zero). Any difference between one's own prediction and the actual frequency is penalized by using their relative entropy (Kullback-Leibler divergence). Thus, prediction score can be high for respondents predicting accurately how others will respond. From the equation, it can be noticed that the information score is independent from the predicted frequency as well as the prediction score is independent of the personal answer. The constant α adjusts the weight given to the prediction score. As α approaches zero, the BTS score relies more to the information score, the game results in "Pareto-dominant expected scores". For α=1, the total scores of the respondents will be ranked in accordance to their accurate prediction for population frequencies, thus the information and the prediction scores are given equal weight.

BTS theorem postulates that a respondent can maximize his or her expected score by telling the truth assuming that everyone else also gives truthful answers - that is a Bayesian Nash Equilibrium (Prelec, 2004). It should be noted that BTS doesn't give high scores to the "most popular answer" – truth telling is incentivized regardless of whether the opinion is rare or common.

The algorithm relies on the assumption that people are Bayesian statisticians (Weaver & Prelec, 2012). People rely on Bayesian inference when constructing predictions about the distribution of the opinions and beliefs of others. They start with a common prior, which is then updated in direction of their own preferences. In accordance with the Bayes rule people with the same preferences will have the same posterior beliefs. Thus, it is assumed that people with the same beliefs will draw the same frequency predictions. What distinguished BTS from other

Bayesian mechanisms is that respondents are not asked about prior or posterior beliefs. Thus, there is no need of pre-estimation of base rates which can produce vague initial knowledge or non-informative priors (Bernardo, 1979). Using the method, complex and subjective questions can be examined even when the objective truth is unknowable.

Previous research

Despite its ease of implementation, little has been done so far to test BTS's validity. However, the results obtained so far are in favor of the approach. For example Weaver & Prelec (2012) tested the method in recognition surveys where it proved to reduce the number of recognized items in contrast with introspection. In those experiments subjects were provided with lists with existent and non-existent items. People in the BTS condition recognized far less non-existent items than those in the control groups. The authors also showed that BTS outperforms a common truth serum – the solemn oath. The solemn oath asks participants to sign a declaration that they will tell the truth during participation. Finally, when used in evaluation of public good, BTS was able to eliminate the hypothetical bias usually caused when using contingent valuation. CV is a common method when assessing the contributions people would make for a certain public good.

Furthermore, John, Loewenstein & Prelec (2012) used BTS to analyze questionable research practices. Motivated by the scientific approach to misconduct research, they sent emails to over 2000 psychological researchers asking about "questionable research practices". They proved that truth-telling incentives in the BTS algorithm tend to induce more truthful answers.

Barrage & Lee (2010) compared BTS with Contingent valuation, "cheap talk" and consequentialism to see which method would correct for the unrealistic overestimation of donations for a public good. Their goal was to find a method which can be used as a tool for eliminating the hypothetical bias which usually occurs in CV experiment settings. However, their results showed that the hypothetical bias is only reduced, but not eliminated when using BTS.

Howie, Wang, & Tsai (2010) created a slight modification of BTS and tested whether the method can be applied to predict the success of a new product. The goal of their research was to test the predictive reliability of the method when at the time of the survey the truth is unknowable: the adoption of a new product. Their results showed that the predictive validity increased when using BTS.

So far the BTS has not been applied in the measurement of WTP and WTA for marketed goods. As people tend to underestimate their WTP and overestimate their WTA, BTS can play a

crucial role when measuring consumers' WTP and WTA for a new product. For marketed goods, correctly estimating WTP and WTA plays a role when predicting the volume of the market and the actual profit (Borges & Knetsch, 1998). Given the importance of accurately predicting market size and gains, truthful elicitation is an important indicator for market dynamics. Moreover, truthful elicitation for marketed goods with substitutes should yield insignificant difference between WTP and WTA. This paper will examine whether BTS will outperform non-incentivized introspection when measuring WTP and WTA for a new product. Thus, the paper will evaluate BTS validity in an additional setting represented by evaluating WTA/WTP for a marketed good. Most importantly, the use of BTS will reveal whether the method is reliable for overcoming the WTA/WTP gap.

The good in hand is a service provided by Uber. To understand what Uber does and what it represents, one should be familiar with the concept of Collaborative Consumption and its main features.

Collaborative Consumption

The development of technology and especially Web 2.0, allowed for the creation of new ways for consumer interaction and behaviour (Perren et al., 2015). The peer-to-peer exchange of goods is now associated with successful companies as Airbnb, Uber, Lyft, Zipcar, Etsy, Kickstarter, all of which have flourished thanks to the Internet. Apart from the fact that the companies work essentially online, another common feature in those business models is the non-ownership implied for utilizing consumer goods and services (Belk, 2013). They are collectively part of the so-called "sharing economy" which includes both for-profit and non-profit companies (Schor, 2014).

Many authors use different terms for that phenomenon such as collaborative consumption, the mesh, commercial sharing business, co-creating, prosumption, product-service systems, access-based communication, consumer participation and online volunteering (Belk, 2013). The most widely used of these is collaborative consumption, defined by Rachel Botsman and Roo Rogers (2011) as follows:

"*The reinvention of traditional market behaviours—renting, lending, swapping, sharing, bartering, gifting—through technology, taking place in ways and on a scale not possible before the internet*".

Belk, 2013 provides an extended definition of CC which includes the role of the facilitator (the company itself which provides web platform) in the peer-to-peer marketplace:

"*Collaborative consumption is people coordinating the acquisition and distribution of a resource for a fee or other compensation*".

Therefore, CC markets alter the traditional buyer-seller marketplace by adding a third party – the facilitator, making the marketplace triadic (Perren et al., 2015).

In such a market the role of the consumer is changing and alternative models of products acquisition and consumption are arising (Bardhi & Eckhardt, 2012). One direction of the shift **is towards temporary access of goods and services rather than ownership** (Bardhi & Eckhardt, 2012; Belk, 2013; Perren et al., 2015). For example, Zipcar (now acquired by Avis) offered its users the opportunity to reserve a car through a mobile application, unlock and drive it with a special membership card they get once a yearly fee is paid (Belk, 2013). Car-sharing services have gained huge popularity among participants and the general public, as they provide alternative means of transportation offering both flexibility for the consumer and environmental gains for the community (Firnkorn & Müller, 2012). The new business model have also spread among traditional car manufacturers as Mercedes, BMW, Volkswagen and Peugeot mainly because young consumers are no longer interested in possessing a car as its maintenance would be both expensive and tiring (Belk, 2013). Moreover, companies recognize the opportunity for expansion since there is an increasing car-sharing demand from cities in an attempt to reduce traffic congestion (Firnkorn & Müller, 2012).

Furthermore, consumers are no longer passively buying goods; **they are becoming actively involved in their production**. By being involved in joint production of value with other consumers and businesses, consumers are becoming "co-creators" or "prosumers" (Denegri-Knott & Zwick, 2011). This trend as observed mainly for websites generating content and sharing information such as Facebook, Twitter, YouTube, eBay, blogs, etc., is expanding to other sectors including transportation, banking, retail and accommodation (Geron, 2013). For example, Uber provides a platform in which consumers can offer taxi-like services. By simply registering and logging on in a mobile application riders and drivers are connected.

An interesting question is **why consumers participate in these new marketplaces**. Research has been scarce in this new domain, however, the main reasons described are saving money and preserving the environment (John, 2013). Hamari, Sjöklint, & Ukkonen (2013) find that people are motivated by factors like sustainability, enjoyment and economic benefits. Their research reveals that sustainability leads to a positive attitude toward CC participation, however personal gains are stronger predictor of actual behaviour. Grenville et al., (2014) describe three main drivers for participation in CC: 1) societal, which encompasses environmental concerns as well as desire for independent lifestyle; 2) economic, which refers to consumers' desire to utilize new sources of income and maximize resource utilization; 3) technological being the presence of facilitators (the web platforms). Additionally, their research find the most common reasons for

sharing to be convenience and price, unique product or service and recommendation (WOM). Schor (2014) lists social, environmental and economic motives as the main drivers. More precisely, the desire to increase social connections and the provision of goods and services at lower costs are the major reasons for participation. These platforms provide opportunities for people to diversify their income, goods and services and "there is potential in this sector for creating new businesses that allocate value more fairly, that are more democratically organized, that reduce eco-footprints, and that can bring people together in new ways" (Schor, 2014).

The focus of this paper will be **Uber** as one of the most well-known examples of CC. Created in 2009 as a private luxury car service, it has now expanded greatly (Cusumano, 2014). Its services are considered innovative in regards to transaction costs and intermediary service presented by their mobile application (Maselli & Giuli, 2015; Rogers, 2015). Uber's main service is a taxi-like experience. What makes it different is the online mobile application which connects riders and drivers, the pre-defined price of a selected route and the cashless transactions. The application enables riders and drivers to easily connect to each other, get price quote on the ride and accept or deny it (Rogers, 2015).

When a rider chooses a route (starting and ending destination), the application calculates how much it would cost. Both riders and drivers are pre-committed to a certain fee. There are no cash payments, all of the transactions are through the passengers' credit cards as a percentage of the fee goes to Uber and the rest to the drivers themselves. By providing the application as a mean of communication and search platform, Uber removes any search costs – there is no need to call a dispatcher and wait. Once a ride is ordered one can see its progress through the application, e.g. the time of the arrival is calculated. All of this is advantageous for passengers who don't like uncertainty in terms of payment and waiting (Rogers, 2015). Uber relies on a "dynamic pricing" strategy, meaning that during rush hours or bad weather, the fees are increased (Cusumano, 2014).

Uber drivers are not regarded as Uber employees, Maselli & Giuli (2015) argue that Uber provides a franchising. The drivers receive the right to use the trademark Uber to provide the service at certain level of standards. At the same time, as not being regarded as employees they have the freedom to choose when to work and make their own schedule. However, drivers do not receive a salary, but as pointed out above a share of a fee. Additionally, visible for everyone is the driver and the rider ratings. Thus, drivers have the flexibility to decide not to accept the rider based on someone's negative review.

What seems to be the result is that Uber is creating a market driven by supply and demand (Rogers, 2015). However, it's not a "free market", because the company is the facilitator which sets the prices.

The researched product in this paper will be an urban delivery service which does not yet exists in the Netherlands. The company's plans to expand its services to urban logistics have already started in certain cities around the world (Badger, 2014). In 2014, it started with UberRush, which is a bike and pad courier service for delivering products and goods from one location to another. The order is placed in the application allowing the customer to easily track it. Recently the company was involved in discussions with over 400 merchants, some of which are Hugo Boss, Tiffany's and Louis Vuitton about using Uber's drivers for delivery purposes (Anderson, 2015). However, Uber changed its business model regarding this application – currently, they are partnering with local businesses in order to improve delivery services. The difference between the initial service and the current one is that UberRush is a "delivery driver", not as before an application where customers can place their orders. That is, the customer directly orders from the business and Uber delivers the product in the background, implying that customers don't really know that their delivery will be handled by Uber until they get the product. The costs for delivery are between $4 and $7, businesses are free to decide whether they want to pay themselves or add the delivery costs to the total order.

Moreover, Uber wants to enter the food delivery business using another Uber application: UberEATS. It will allow customers to get delivered dishes from different restaurants quickly for a flat fee of $3 or $4 in the US. The application is available for several cities in the US and in Paris. The company is again partnering with local restaurants and it delivers meals at different times. Along with the cost of the meal, the customer has to pay a flat delivery fee.

As this may be an opportunity to diversify their products and engage even more drivers, concerns have also arisen mainly for the work arrangements (Anderson, 2015). An online discussion[3] has revealed payment concerns regarding people's willingness to accept to participate and willingness to pay to get the service. For example:

- Nikki Baird, managing partner at RSR Research posed the questions:

"How do the drivers feel about a $3-4 flat rate delivery? Is that reasonable for them, given that Uber covers nothing related to car costs?"

- Kai Clarke a CEO from American Retail Consultants stated:

"No. Uber is a great idea, but hailing a taxi (which is what Uber really is) is not necessarily cost, or service effective. Who wants to pay $4 for a pizza delivery when the pizza costs $5? Or a sandwich that costs $6? How will Uber compete against Amazon and others who are already ramping up in this arena for less? My $2.95 delivery from Amazon is great!"

[3] For more information see http://www.retailwire.com/discussion/18251/uber-plans-to-deliver-everything

Keeping in mind these worries and the fact that Uber's success is determined by its drivers and customers, it is reasonable to explore what the true valuations of WTP and WTA will be so that we can predict what the market for food and retail delivery will look like for Uber. As a price-setter, Uber is responsible for determining fair prices for both its drivers and customers. The recent 2015 1099 Economy Workforce Report concluded that people working for on-demand services such as Uber are not satisfied with their experience so far. The study found that high attrition rate occurs mainly due to insufficient pay: 42.1% of the respondents indicated this as a factor for abandoning this type of work. Moreover 26.4% indicated the "insufficient flexibly in schedule" and 15.9% "inconvenient or inflexible location" as main factors. Unfair actions from a company might result in customers' and employees' willingness to "punish unfair transactions" (Kahneman, Knetsch, & Thaler, 1986). That is, unfair actions from a company whose goal is to "exploit unfair profit opportunities" might cause reluctance to participate in such transactions which poses a major challenge for a given company in competitive markets. Hence, it is important for Uber to obtain users insights for its service evaluations which can assist in determining a fair pricing policy, otherwise consumers as well as drivers might be tempted to switch to competitors such as Lyft, Curb and Sidecar.

As mentioned earlier, the relevance of the WTP-WTA gap for markets concerns volume trade and actual profit. Given the fact the people value losses more than gains (Kahneman et al., 1991), setting higher selling price and obtaining lower buying value for the same item would imply less transactions than predicted by the standard assumption that losses and gains are valued identically (Borges & Knetsch, 1998). That is, fewer buyers will be willing to obtain a good for higher price than they are willing to pay. Consequently, fewer gains will be obtained from the transactions. Therefore, it's essential for companies to correctly predict supply and demand for their products and services for determining market size and therefore, company's market share.

Finally, establishing the true ratio for WTA and WTP can point out whether CC platforms actually are sustainable. Rachel Botsman and Roo Rogers (2011) argue that CC has the potential of creating sustainable consumer behavior, which in the long-term will be beneficial for the individual, the society and the environment. However, there are many challenges concerning this new form of marketplaces. Its virtues are being heavily debated as the new businesses raise attention to legal and regulatory issues (Perren et al., 2015). The lack of well-establish regulations poses legal challenges for the newly established businesses, despite their wide public acceptance. Apart from the legal issues, which are not subject of the present report, problems and critiques regarding users' satisfaction are arising. Uber's critics so far are in six directions: 1) unfair competition with taxi drivers; 2) by vertical and horizontal integration, it seeks to become a monopoly; 3) providing unsafe services; 4) invading customer's privacy; 5) enabling

discrimination and 6) undermining working standards and low compensation (Rogers, 2015). All of these raise the question of whether the sharing economy is sustainable after all.

Estimating the true WTP and WTA ratio with the aid of BTS can help move towards better understanding of CC dynamics. The paper will evaluate whether BTS can be used as a tool for measurement WTP and WTA and therefore, correctly estimating market volume and profits. Additionally, it will contribute to the scarce literature on the sharing economy by providing insights about the potential demand and supply in the urban logistics business.

Hypothesis

The research will aim to find if applying the BTS in a context of the sharing economy would reduce the ratio between individuals' willingness-to-pay and willingness-to-accept. The main purpose is to examine BTS validity in a context where it hasn't been tested before and compare it with the traditional method, introspection, for deriving consumers' WTP and WTA. The main hypotheses that will be tested in this research are as follows:

H1. The difference between the mean values of WTP and WTA will be insignificant when the BTS is used as an incentive-compatible method.
H2. The difference between the mean values of WTP and WTA will be significant when no incentive-compatible method is applied.

Research Methodology and Data

Scoring Mechanism

A common approach of estimating WTP and WTA involves asking respondents to state their valuations for the product in hand, e.g. providing a continuous signal. The truth scoring of BTS is suitable for questions with *k* answers, however, for questions with continuous signals it can be rather inconvenient.

Typically, to compute the final payout of a respondent *r* the BTS mechanism implies that:

1) Every subject *r* is asked to provide answers to two questions:
   - Personal opinion x_k^r, e.g. endorsement of answer k;
   - Frequency of people endorsing answer k y_k^r.
2) The average $\bar{x}_k$ of the answers and the geometric mean $\bar{y}_k$ of the predictions are computed as follows:

$$\bar{x}_{k=}\frac{1}{n}\sum_{r=1}^{n} x_k^r \ , \qquad log\bar{y}_k = \frac{1}{n}\sum_{r=1}^{n} logy_k$$

   - The final payout for every respondent r is estimated by means of the following formula:

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

However, when exploiting this approach with continuous signals, an inconsistency between the calculation of the information scores and the prediction scores can occur. The first question asks respondents to provide their valuation which is stored as a continuous variable. In a typical BTS setting, the second question will request the respondents to provide a prediction of the distribution of people willing to pay/accept the same price as they do. However, asking respondents to predict the distribution of answered spread across multiple categories (e.g. 0 to 10, 0 to 50, 0 to 200, etc.) is not very handy and convenient. Thus, a problem estimating the predictions' scores distribution occurs. Moreover, the information and the prediction score will represent different dimensions, namely, continuous probability distribution for the endorsement frequencies and discrete probability distribution for the predicted frequencies. The differences in the distributions will cause complication estimating the final scores as they rely both on the information score and the prediction score, which depict different dimensions.

To handle this issue, a slight alternation of the BTS approach and computation is proposed. The first question remains the same, that is - respondents are asked to provide their valuations

which are recorded as a continuous variable. The second question, however, is modified. Instead of stating the frequency prediction of their endorsement, subjects are asked to provide the average WTA or WTP of other respondents, which is recorded again as a continuous signal. To compute the final payoff, both valuations are assigned into intervals, ensuring the consistency between both scores. The personal valuations are assigned into categories, assuming 1 euro precision. The predicted frequencies distributions for each respondent are estimated as Poisson distribution is assumed, using respondents' *r* answer as a mean. The average WTA/WTP are categorized by the means of Poisson distribution. Poisson distribution is adopted because of the following reasons: 1) it is frequently used for categorical data; 2) it is a good approximation of the underlying distribution even when nothing else is known about the data; and 3) using just the mean, an estimation of the distribution is possible (Kianifard & Zelterman, 2000). The result of the additional computations is a probability distribution for each respondent which can be interpreted as estimated frequencies. The final payoff can be computed by means of the following procedure:

1) Every subject *r* is asked to provide answers to two questions:
   - Personal valuation of the product $x^r = (x^r_{1,\ldots\ldots} x^r_k)$;
   - Estimation of the average valuation $y^r = (y^r_{1,\ldots\ldots} y^r_k)$, using each subject's prediction of the average mean to calculate his/her own Poisson distribution.
2) Compute of the average $\bar{x}_k$ of the answers and the geometric mean $\bar{y}_k$ of the predictions:

$$\bar{x}_k = \frac{1}{n}\sum_{r=1}^{n} x^r_k$$

$$log\bar{y}_k = \frac{1}{n}\sum_{r=1}^{n} y_k$$

Compute the final payout for every respondent r:

$$BTS\ score = \sum_k x^r_k\, log\frac{\bar{x}_k}{\bar{y}_k} + \alpha \sum_k \bar{x}_k log\frac{y^r_k}{\bar{x}_k}$$

Experimental design and procedure

A computer-administered survey was executed which included instructions followed by three main sections. The first part was of main interest to this research and examined subjects' valuation for Uber's new product. Participants were requested to read a short description of Uber's new product, followed by questions regarding their valuations of WTP or WTA in an open-

ended manner. The second part of the questionnaire included basic demographic questions. Finally, subjects were asked to leave their email addresses in case they win a prize of 15 Euros.

The survey differed in instructions in regards to the incentive-compatible method and the control group and questions on willingness to pay or accept, thus, a 2x2 between-subject design was created: truth-telling incentive (BTS or control) and WTP-WTA condition. The questionnaires were distributed online, where the four different conditions were randomized, such that each respondent had an equal chance of being in a given condition. In the treatment condition, the first section included two questions regarding WTP or WTA in accordance with the BTS. Subjects were asked to first provide their own valuation of the product and then an estimate of the average valuation that other participants would make. In the control condition, respondents were only asked to provide their own valuations.

The instructions for the BTS condition were adopted from Weaver & Prelec (2012). Participants were not given a detailed description of how the scoring mechanism works, however, they were told that the truthful responses are rewarded even though the accuracy of their valuation is not known. It implied that their best strategy is to give truthful answers (the survey design can be found in the Appendix A. Survey Design). Participants were informed that by acquiring high scores, they have a chance of winning 15 Euros. In the control group, participants were told that one person can win 15 Euros at random and they were asked to answer carefully and honestly.

Subjects

210 respondents received and opened the questionnaire. Of them 83 responses were incomplete and 27 respondents were pre-screened based on their knowledge of Uber. Participants not familiar with Uber as a service provider were now allowed to continue and finish the questionnaire, they were briefly thanked for their participation. The reason for that choice was that in case where respondents' lack information regarding the product in hand, they might engage in "satisficing" (Krosnick, 1991). That is, subjects might provide a satisfactory answer instead of a truthful one if they are not familiar and lack information on the topic asked. Overall, 100 valid responses were collected.

The descriptive analysis below shows the demographics of the 100 respondents regarding age, gender, educational background and place of residence (see Table 1). The mean age is 25.5 across the sample and 44% of the respondents were males and 54% - females. 62 of the respondents have never used Uber and 38 stated that they have used it. Similar proportions are observed across the conditions.

Table 1, Demographics

TOTAL (n=100)						
Age (mean)	Gender	%	Education	%	Residence	%
25.5	Male	44%	High-school	11%	Netherlands	80%
	Female	54%	MO/HBO	8%	France	2%
	Missing	*2%*	Bachelor	40%	UK	1%
			Master	35%	Bulgaria	6%
			Doctorate	1%	Germany	3%
			Missing	*5%*	Japan	1%
					Italy	1%
					Poland	1%
					Missing	*5%*
WTA Control group (n=32)						
Age (mean)	Gender	%	Education	%	Residence	%
26	Male	50%	High-school	9%	Netherlands	75%
	Female	50%	MO/HBO	3%	France	3%
			Bachelor	50%	Bulgaria	9%
			Master	31%	Germany	9%
			Doctorate	3%	*Missing*	*3%*
			Missing	*3%*		
WTP Control group (n=24)						
Age (mean)	Gender	%	Education	%	Residence	%
25	Male	42%	High-school	13%	Netherlands	92%
	Female	58%	Bachelor	38%	Poland	4%
			Master	46%	*Missing*	*4%*
			Missing	*4%*		
WTA Bayesian Truth Serum group (n=20)						
Age (mean)	Gender	%	Education	%	Residence	%
24.6	Male	40%	High-school	15%	Netherlands	75%
	Female	50%	MO/HBO	20%	France	5%
	Missing	*10%*	Bachelor	30%	Bulgaria	5%
			Master	25%	Japan	5%
			Missing	*10%*	*Missing*	*10%*
WTP Bayesian Truth Serum group (n=24)						
Age (mean)	Gender	%	Education	%	Residence	%
25	Male	42%	High-school	8%	Netherlands	88%
	Female	58%	MO/HBO	13%	UK	4%
			Bachelor	38%	Bulgaria	4%
			Master	38%	Italy	4%
			Missing	*4%*	*Missing*	*4%*

Incentives

In the control group, as stated in the instructions, a winner was chosen at random and rewarded 15 Euros. The person was contacted by the email provided in the last part of the questionnaire.

In the BTS group, the respondents' total scores were calculated in accordance with the scoring mechanism described above.

1) The personal valuations were assigned into 11 categories (intervals) with 1 euro precision between each one. The intervals are between 0 and 10. The last interval (10) contains valuations which are exactly or more than 10.
2) Based on the intervals the frequency of each valuation (price) and its actual percentage were calculated:

$$\bar{x}_{k=}\frac{1}{n}\sum_{r=1}^{n}x_k^r$$

3) Using respondents' average WTP or WTA as a mean and the intervals created in the first step, for each respondent Poisson distribution was computed:

$$P(n;\ \mu) = \frac{(e^{-\mu})(\mu^n)}{n!}$$

where μ is the average WTP or WTA each respondent stated, and

n represents each interval.

4) The geometrical mean for the estimated frequencies was calculated based on the Poisson distribution:

$$log\bar{y}_k = \frac{1}{n}\sum_{r=1}^{n}y_k$$

5) The Information score was calculated based on the geometric mean and the actual average:

$$Information\ score = \sum_k x_k^r\, log\frac{\bar{x}_k}{\bar{y}_k}$$

6) The prediction score representing the "penalty proportion to the relative entropy between the empirical distribution and r's prediction of that distribution" was calculated as suggested by (Prelec, 2004):

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

7) Finally, the sum of the prediction score and the information score resulted in the total BTS score calculated for each respondent:

$$BTS\ score = \sum_k x_k^r\ log \frac{\bar{x}_k}{\bar{y}_k} + \ \alpha \sum_k \bar{x}_k log \frac{y_k^r}{\bar{x}_k}$$

In this way, respondents' total scores were calculated. Unlike the control group, where a respondent was chosen at random, for the BTS condition, the respondent with the highest score was rewarded 15 Euros. He was contacted by the email provided in the last part of the questionnaire.

Results

Analysis

Analysis of the data was done using the Mann-Whitney U test and Bayes factors. To test for robustness the tests were performed twice, the second time accounting for outliers.

Figure 1 presents the mean valuations of the new service from Uber in each condition. The WTA in the control group is substantially higher than the WTP. On average, people stated that they would be willing to accept an amount 30% higher than the one they are willing to pay for the same service. Results from the BTS conditions represent a different picture. In this condition, on average people expect to accept and pay the same amount. Interestingly enough, these results suggest that people are willing to pay a little bit more for the service than they are willing to accept for it.

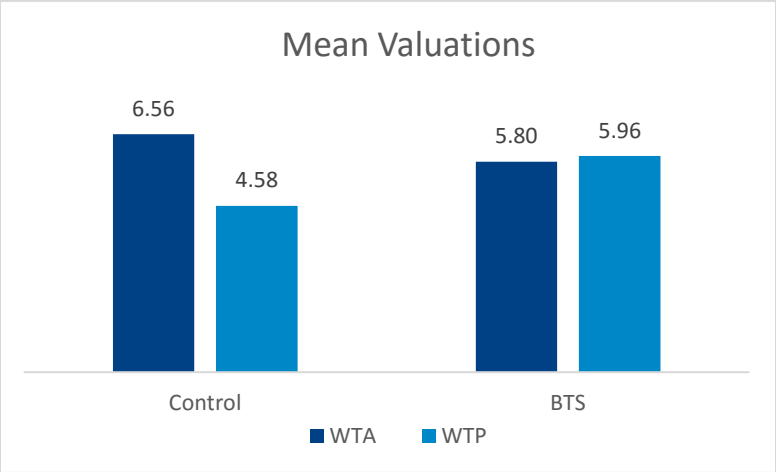

Figure 1, Mean Valuations

The data consists of 2x2 independent samples with a between-subjects design, thus, for estimating whether there is a significant difference between the mean valuations two tests were considered, namely, the Independent sample t-test and Mann-Whitney U test. Using the Independent sample t-test would imply that the data fit a normal distribution. If the normality assumption does not hold, the data interpretations might not be "reliable and valid" (Razali & Wah, 2011). To test for normality, two tests were conducted with the aid of SPSS – the Kolmogorov-Smirnov and Shapiro-Wilk tests. The null hypothesis of both tests states that the data follows a specified distribution. From *Table 2* it is evident that for all groups in both conditions the null hypothesis of a normal distribution is rejected at significance level $p<0.001$. The data does not fit a normal distribution and continuing the analysis using the Independent samples t-test would be inappropriate since the assumption of normality is not met.

Table 2, Normality tests

	Kolmogorov-Smirnov	*Sig.*	*Shapiro-Wilk*	*Sig.*
BTS-WTA	.262	*.001*	.682	*.000*
BTS-WTP	.449	*.000*	.379	*.000*
Control-WTA	.219	*.000*	.898	*.005*
Control-WTA	.285	*.000*	.798	*.000*

In such a case, the Mann-Whitney U test is suitable since it is a non-parametric test which does not require the data to fit a normal distribution (Mann & Whitney, 1947). The collected data comes from a small sample of subjects and the test is appropriate in such scenario to derive reliable and valid conclusions (Nachar, 2008). Additionally, in the social sciences, the test is one of the most frequently used non-parametric tests (Nachar, 2008).

Mann-Whitney U test

The null hypothesis of the Mann-Whitney U test states that the two groups of interest come from the same population meaning that they have the same distribution (H0: $\mu 1 = \mu 2$). The alternative hypothesis claims that the distribution of the first group differs from the distribution of the second one (H1: $\mu 1 \neq \mu 2$). In other words, the test investigates whether there is a significant difference between the means of the two groups.

Assumptions of the test

There are three assumptions that should be met when performing the Mann-Whitney U test (Nachar, 2008):

1) The two groups must be randomly drawn from the population.
2) There is independence of observations, implying that there is independence between the observations in each group and between groups. In other words, each data point should be from a different participant.
3) Data should be at least at ordinal level.

The three assumptions of the test were met as:

1) Participants were randomly assigned to the different conditions when they opened the questionnaire.
2) Participants were allowed to fill out only one questionnaire to which they were assigned randomly, thus the design of the study was between-subjects, implying that there was independence between and within groups.
3) The data is at ratio scale. Participants were asked to fill in their valuations for the WTP and WTA in open-ended questions and continuous signals were collected.

Mann-Whitney U method

The test compares the two independent samples to each other. Comparing whether the sum of the ranks of the two samples are similar, the tests calculates a U-statistics to test if the two samples come from the same population.

Table 3, Mann-Whitney U test

	n	*Sum or Ranks*	*Mann-Whitney U*	*Sig.*
BTS	***44***		***160.000***	***.054***
WTA	20	530		
WTP	24	460		
Control	***56***		***294.5***	***0.132***
WTA	32	1001.5		
WTP	24	594.5		

Looking at the results in *Table 3* it can be concluded that the WTA is not significantly different from the WTP in both the BTS and the Control group ($p>0.05$). In other words, for both conditions, we **cannot reject the null hypothesis** and can conclude that there are not significant differences between the means.

The conclusion based on the Mann-Whitney U test implies that researchers should not engage in using the BTS to enhance data quality as it does not lead to better predictions compared to the conventionally used method. These conclusions could result from a lack of statistical power. That is, it could have been possible that the design of the study failed to detect a significant difference between the two groups because it was simply too small. The sample size affects the power and it might have not be sufficient to achieve a power of at least 80%. Thus, the probability of rejecting a false null hypothesis increases when the statistical power decreases. Moreover, the typical interpretation of significance tests suggests that they "may be used only to reject hypotheses and do not offer an assessment of the strength of the evidence in favour of the null hypothesis" (Kass & Raftery, 1995). Since in both conditions, the alternative hypothesis was rejected, it was of interest to further test to what extent in both conditions the null hypothesis is supported using a Bayesian t-test.

Bayes Factor

The Bayes factor could answer the question how strongly the data supports the null hypothesis for the BTS and control condition over the alternative. It represents an alternative of the traditional t-test that "allows researchers to express a preference for either the null hypothesis or the alternative" (Rouder, Speckman, Sun, Morey, & Iverson, 2009). It can provide insights of how strongly the data supports one hypothesis (H0) over the other (H1). Rouder et al. (2009) argue that the conventional significance tests can lead to 1) inability for the researchers to state an evidence for the null hypothesis and 2) tendency to overstate the alternative hypothesis. For example, as mentioned above given a non-significant p-value one cannot conclude whether there is or there isn't evidence for the null hypothesis.

The Bayes factor is comparative and it can help analysts determine the magnitude of the difference between the two samples. It essentially calculates the ratio of the likelihood of the data under each of the hypotheses:

$$BF = \frac{Pr(H_0|data)}{Pr(H_1|data)}$$

The result is interpreted directly. For example, if BF is equal to 10, the null hypothesis is 10 times more likely than the alternative given the data. Thus, an increase of the BF indicated an increase in the support for the null hypothesis. *Table 4* represents the conventionally used scale for BF's interpretations (Jeffreys, 1961).

Table 4, Bayes Factor

Bayes Factor	*Evidence*
1-3	Anecdotal
>3	Substantial evidence
>10	Strong evidence
>30	Very strong evidence
>100	Decisive

The calculation of the Bayes factor was done with the aid of the R package (see Appendix B. R-code). For the control group (BF=1.13), the value of 1.13 implies that there is no evidence neither for the null hypothesis nor for the alternative. For the BTS condition (BF=3.34), however, there is **substantial** evidence that the null hypothesis of equality is three times more probable than the alternative given the data. These results suggest that although in both conditions there is not a significant difference between the means, equality of WTP and WTA more probable in the BTS condition than the control group. Thus, using BTS to explore people's willingness to pay and willingness to accept is justified as it provides more truthful and significant results. So far the study has managed to prove hypothesis 1, namely **"**The difference between the mean values of WTP and WTA will be insignificant when the BTS is used as an incentive-compatible method. ", but has not given a definitive answer to whether the difference in the mean values of the WTP and WTA in the control group was significant.

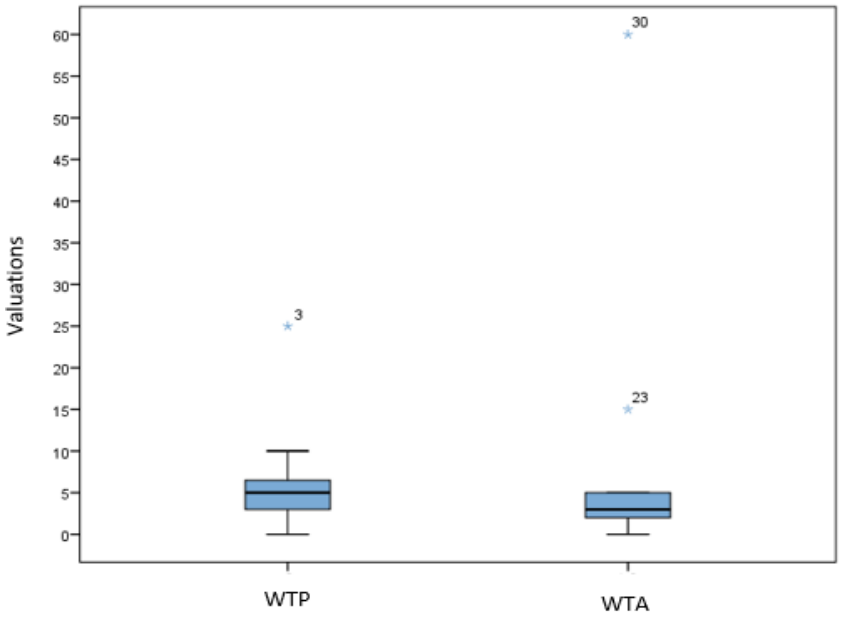

Figure 2, BTS outliers

Outliers

Since the results indicated that in the BTS condition on average, people are willing to pay a bit more than the amount they are willing to accept, the data was examined further. It was tested for outliers and the same tests as described above were performed excluding only the extreme outliers. Extreme outliers are those data points whose values are 3 times the interquartile range

(the middle 50 values). That is, an extreme outlier would be located about 3 times above or below the middle 50 values. Looking at *Figure 2* and *Figure 3*, it can be seen that in both conditions there are outliers. In the BTS condition, there are 3 "extreme" ones while in the control group there is only one "extreme" and three "out values". With the aid of both graphs, it is easy to observe the differences between the two groups. Despite the fact that in the BTS group there are three extreme outliers, it can be seen that the valuations for both WTP and WTA are more compressed toward the means. In the control group, the valuations of WTP are more spread between 0 and 20, while the prices indicated for the WTA are less spread.

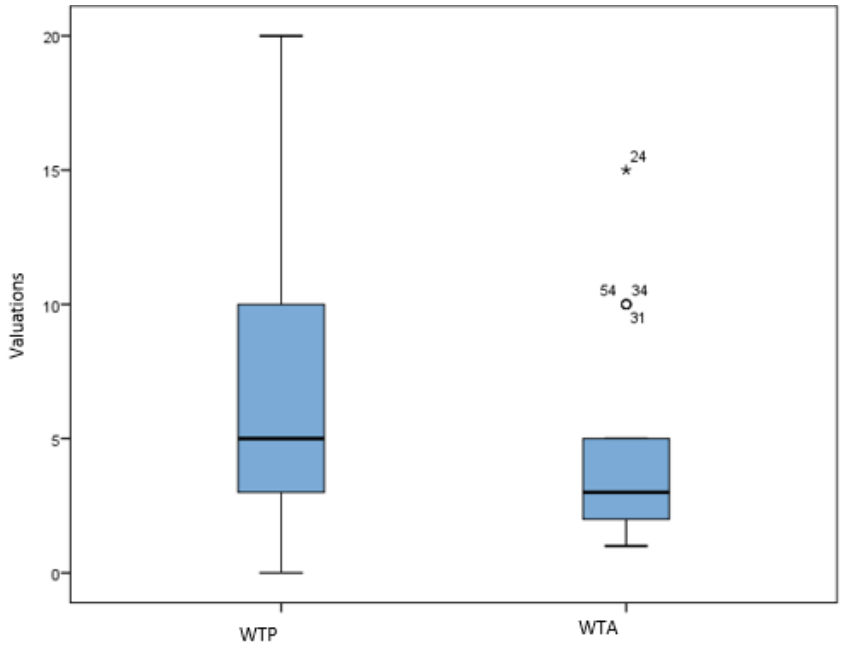

Figure 3, Control outliers

After excluding the outliers from the data differences in the outcomes are observed. First, the mean valuations for both groups changes (see *Figure 4*). In this case, a reversed picture is observed. As the outliers are removed, for both group the mean valuations are decreased. In the BTS condition, this time, the WTA is about 35% higher than the WTP. For the control group, the difference is smaller (approximately 18%), yet still existent.

To further check for significant difference, both the Mann-Whitney U test and the Bayes Factor were estimated. The conclusions based on the Mann-Whitney U test change as well. *Table 5* indicates that in the BTS group the null hypothesis is rejected at $p<0.05$) implying that the mean differences between the two samples are significant. For the control group the conclusions do not change – there

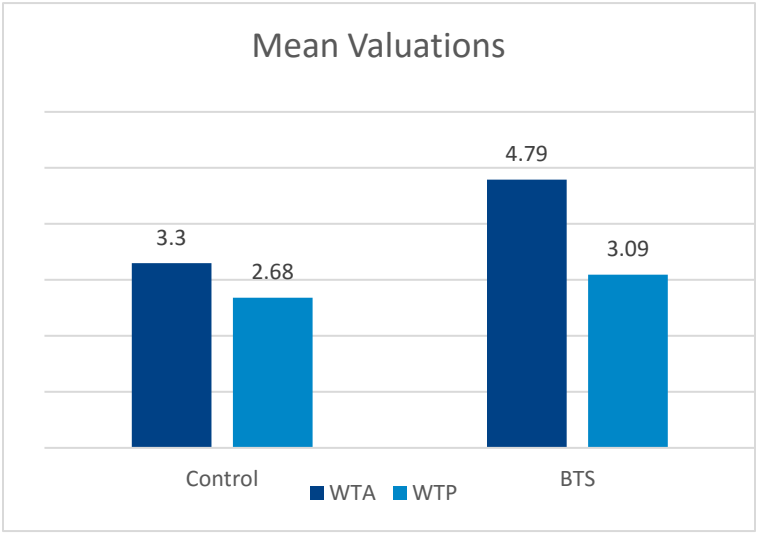

Figure 4, Mean valuations without outliers

is no significant difference between the two samples (p>0.05), even when the outliers are excluded from the data.

Table 5, Mann-Whitney U excluding Outliers

	n	*Sum or Ranks*	*Mann-Whitney U*	*Sig.*
BTS	***42***		***140***	***0.042***
WTA	19	487		
WTP	23	416		
Control	***55***		***265***	***0.074***
WTA	32	999		
WTP	23	541		

Finally, looking at the Bayes factor for the BTS condition (BF=0.25), there is no evidence of support for the null hypothesis. However, the alternative hypothesis is 3.95 times more probable than the null, considering the data. This suggests that there is substantial evidence to accept the alternative hypothesis meaning that **there is a significant difference** between WTP and WTA valuations. For the control group (BF=0.55) again there is no evidence neither for the null hypothesis nor for the alternative. Removing the outliers from the data shows that under the BTS condition there is a notable difference between the two valuations respondents provided, yet there is no conclusive evidence for either of the hypotheses in the control group.

Removing the outliers from the BTS sample changes the direction of the conclusions greatly. It revealed a completely different picture from before, signalling the need for further analysis with an increased sample size. The outliers changed notably the results and suspected is that the main reason is the low number of participants. Since only 44 respondents participated in the BTS condition, the results greatly change when removing 3 of them. As the results showed that the BTS condition is highly sensitive to outliers, it indicates that the method is inferior. For that reason, future research recruiting more participants is recommended to replicate the analysis.

Conclusions

General Discussion

The main goal of this paper was to explore whether a relatively new incentive-compatible method, the Bayesian truth serum, would reduce the ratio of the willingness to pay and willingness to accept for a new product by Uber compared with the traditionally used non-incentivized introspection. My proposition was that adopting BTS over non-incentivized methods would greatly reduce the discrepancy between WTP and WTA. The goal of the paper was to examine whether BTS would prove itself to be a more truthful elicitation technique in a setting where truthful answers cannot be verified at the time of the study. The context for testing the method was the sharing economy. More precisely, respondents were ask to give their WTP and WTA valuations for a new product created by Uber, a company adopting a shared economy business model.

The study used the BTS as an incentive-compatible method in a context in which it hadn't been tested before. Due to the nature of the study and the data collection process itself, an altered version of the method in regards to its scoring mechanism was proposed. The BTS works in a simple manner – it awards truthful responses by assigning them high truth score. The truth score is calculated based on a personal response as well as respondents' prediction about the distribution of the answers of others. The typical setting in which answers are collected is through questions with *k* answers, however, the application of the method under answers with continuous signals is absent. Thus, an adapted version was proposed within this research, which accounts for the consistency in the computational procedures regarding respondents' personal opinion and their prediction of the frequency of the answers.

The data analysis was done using two tests, namely the Mann Whitney-U test and Bayes factor. The Mann Whitney-U test was used to assess whether there is a statistically significant difference between the mean valuations of WTP and WTA in both conditions. The results suggested that for both the BTS and the control group, there are no significant differences in the means. These results indicate that the BTS's usage is not justified as it produces the same conclusions as the non-incentivized method. This, however, does not indicate whether the null hypothesis should be accepted and how strongly the data support it. Thus, the Bayes factor was applied, which helps identify how strongly one hypothesis is supported over the other, given the data. The results in this case indicated that for the control group there is no evidence of whether the data supports the null or the alternative hypothesis. For the BTS conditions, the results indicated otherwise. There is substantial evidence that the data supports the null hypothesis over the alternative. Such a result would indicate that using BTS over traditionally employed non-incentive compatible method yield more accurate data, justifying its adoption. It further can be

argued that this condition triggered respondents to provide truthful answers and eliminated the gap between WTP and WTA in a case where truthful answers are unobservable. The strong evidence of support for the BTS, signalled that the method is more reliable than the traditional non-incentivized technique and researchers should consider adopting it.

To further test for robustness, the data was also checked for outliers and the same tests were performed again without the extreme outliers. In the control group, the conclusions derived from the Mann-Whitney U test and the Bayes factor did not change. However, the findings regarding the BTS conditions changed significantly. First, it was noted that the Mann-Whitney U test indicated significant difference between the two means, rejecting the equality between the two means. Second, applying the Bayes factor, it was found that difference between the two means is almost **4 times** more probable than the equality, giving substantial evidence for the fact that there is in fact a significant difference between the means of the two groups. The outliers have seriously affected the data. This can be explained by the fact that the sample size for both conditions was not high enough. 44 people participated in the BTS condition and removing 3 of them, was shown to significantly change the direction of the results. The BTS eliminates the difference between WTP and WTA only when the outliers are present. Thus, the results are interpreted as finding **no support** in using the BTS for measuring WTP and WTA due to its sensitivity to the outliers. Further research is needed to provide a clearer overview of whether the method should be applied in that domain.

Contributions & Implications

The study showed that BTS performs well in a setting where the truthful answers are unknowable only when outliers are present in the data, however, removing outliers from the sample was shown to significantly affect the data, completely changing the results. The research adds up methodologically to the emerging stream of literature on the Bayesian truth serum by presenting a novel method of how to handle continuous data. There are mixed conclusions regarding BTS validity (e.g. Weaver & Prelec (2012); Barrage & Lee (2008)), and this paper does not support the notion that the method's usage is justified, however, recommends further investigation with a bigger sample. The paper also adds to the literature stream of the WTA/WTP gap. More specifically, the research contributes to the vast amount of studies exploring the experimental design and its incentive-compatibility as a way to reduce the ratio by applying a relatively new methodology that has not been tested in that area. Similarly, to Horowitz & McConnell (2002) review of WTA/WTP studies, the results indicate that using incentive-compatible methodologies actually yields higher ratios than introspection questioning their justification.

Exploring whether BTS or introspection should be used to derive WTP and WTA for new products, not yet placed on the market, the paper disregards both methodologies. The introspection does not support neither of the hypothesis and the BTS data is highly vulnerable to the outliers, questioning its robustness. As estimating product volume and profit is of major concern to businesses as wrong valuations could lead to fewer gains and less volume trade than initially assumed (Borges & Knetsch, 1998), future research of whether business can adopt BTS as a tool for correctly measuring WTP and WTA for accurate estimation of market volume, production and gains is needed.

As this research was administered in the context of a company being part of the sharing economy, the paper contributes to the scarce literature in that domain. There is an ongoing debate of whether the sharing economy is sustainable and whether companies as Uber are fair price setters (Botsman and Roo Rogers, 2011; Perren, Administration, & Florida, 2015; Rogers, 2015). One of the main characteristics in a marketplace representing collaborative consumption is that a person can be both a seller and a buyer, and establishing a fair price policy is crucial for the company. As the study explored a new product for Uber, examining what the true valuations of WTP and WTA can be helpful for predicting what the market will look like. The findings of the current study, however, does not suggest that on average people are willing to pay as much as they are willing to accept. Given these results and the findings from the 2015 1099 Economy Workforce Report, indicating that the high attrition rates for Uber drivers are due to insufficient pay, the company can face problems concerning fair earnings. That is, if customers and employees of the company feel that the company is being unfair to them in order to exploit profit opportunities, they might punish it, e.g. by switching to competitors.

Limitations & Future Research

One limitation is that the sample size was limited and the data is highly affected by each data point. This was shown by removing the outliers in the data and performing the analysis without them. Although the results for the control group did not change, the BTS condition was highly affected by the outliers and the directions of the conclusions changed. Future research should replicate the experiment and involve more respondents in both conditions. In such a way it could provide a clearer answer of whether incentive-compatibility is justified and whether BTS is a more accurate tool over introspection in exploring WTP and WTA.

Furthermore, the study was conducted such that it asked for consumers' valuations without imposing any restrictions. That is, respondents were required to fill in any possible amounts that they would pay/accept for the service provided by Uber. That led to the appearance of extreme outliers (e.g. a person has stated that he'd pay 60 Euros to use this new product). Future research can replicate the experiment, but restrict respondents' valuations in a given

range. Such an approach would also allow the researcher to collect additional data on a minimum and maximum value that can be used to better estimate the underlying distribution for the frequencies' prediction.

Another limitation is that incentives were not given directly and to everyone. Participants were asked to provide their emails so that they can be contacted in case they win the prize of 15 Euros. However, a more plausible and engaging way to conduct the survey, would be for participants to be scored directly and know and receive the amount of their rewards instantly. Moreover, all of the participants were required to provide their own opinion and a prediction about the distribution of answers, increasing respondents' burden. Researchers should consider using a small fractions of the respondents to derive the prediction regarding answer's distribution and use those predictions as a benchmark to calculate all participants' truth scores. As Weaver & Prelec (2012) note:

"*From a theoretical standpoint, it would be sufficient to elicit predictions from a small number of randomly selected respondents and use their predictions to calculate initial iscores. The remaining respondents would only be required to provide answers, as in a traditional survey. The iscores could be periodically updated as the data on empirical proportions accumulates. An advantage of having provisional iscores in hand is that respondents could be scored and rewarded as soon as they enter a response*"

In this way, respondents' burden will be reduced and it can be helpful to score and reward participants immediately (Weaver & Prelec, 2012).

References

Bardhi, F., & Eckhardt, G. M. (2012). Access-Based Consumption: The Case of Car Sharing. *Journal of Consumer Research*, *39*(4), 881–898. doi:10.1086/666376

Barrage, L., & Lee, M. S. (2008). A Penny for Your Thoughts : Incentivizing Truth in Stated Preference Elicitation, (February), 1–21.

Becker, G., DeGroot, M., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, (1), 226–232. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/bs.3830090304/full

Belk, R. (2013). You are what you can access: Sharing and collaborative consumption online. *Journal of Business Research*, *67*(8), 1595–1600. doi:10.1016/j.jbusres.2013.10.001

Berry, J., Fischer, G., & Guiteras, R. (2011). Incentive Compatibility in the Field: A Test of the Becker-Degroot-Marschak Mechanism. Retrieved from http://www.ncsu.edu/cenrep/workshops/TREE/documents/Guiteras_April2011.pdf

Borges, B., & Knetsch, J. (1998). Tests of market outcomes with asymmetric valuations of gains and losses: Smaller gains, fewer trades, and less value. *Journal of Economic Behavior & Organization*, *33*, 185–193. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167268197000905

Brookshire, D., & Coursey, D. (1987). Measuring the value of a public good: an empirical comparison of elicitation procedures. *The American Economic Review*, *77*(4), 554–566. Retrieved from http://www.jstor.org/stable/1814530

Brown, T. C. (2005). Loss aversion without the endowment effect, and other explanations for the WTA–WTP disparity. *Journal of Economic Behavior & Organization*, *57*(3), 367–379. doi:10.1016/j.jebo.2003.10.010

Brown, T. C., & Gregory, R. (1999). Why the WTA–WTP disparity matters. *Ecological Economics*, *28*(3), 323–335. doi:10.1016/S0921-8009(98)00050-0

Carman, K. G., & Kooreman, P. (2011). Perception of Probabilities Flu Shots , Mammograms , and the Perception of Probabilities, (5739).

Carson, R. T., & Groves, T. (2007). Incentive and informational properties of preference questions. *Environmental and Resource Economics*, *37*(1), 181–210. doi:10.1007/s10640-007-9124-5

Coursey, D., Hovis, J., & Schulze, W. (1987). The disparity between willingness to accept and willingness to pay measures of value. *The Quarterly Journal of Economics*, *102*(3), 679–690. Retrieved from http://ajae.oxfordjournals.org/content/72/4/local/advertising.pdf

Cusumano, M. a. (2014). How traditional firms must compete in the sharing economy. *Communications of the ACM*, *58*(1), 32–34. doi:10.1145/2688487

Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology*, *25*(1), 1–17. doi:10.1016/0022-1031(89)90036-X

Denegri-Knott, J., & Zwick, D. (2011). Tracking Prosumption Work on eBay: Reproduction of Desire and the Challenge of Slow Re-McDonaldization. *American Behavioral Scientist*, *56*(4), 439–458. doi:10.1177/0002764211429360

Firnkorn, J., & Müller, M. (2012). Selling Mobility instead of Cars: New Business Strategies of Automakers and the Impact on Private Vehicle Holding. *Business Strategy and the Environment*, *21*(4), 264–280. doi:10.1002/bse.738

Friedman, D., & Massaro, D. W. (1998). Understanding variability in binary and continuous choice. *Psychonomic Bulletin & Review*, *5*(3), 370–389. doi:10.3758/BF03208814

Grenville, A., Samuel, A., & Owyang, J. (2014). SHARING IS THE NEW BUYING HOW TO WIN IN THE COLLABORATIVE ECONOMY.

Guiso, L., Jappelli, T., & Terlizzese, D. (1992). Earnings uncertainty and precautionary saving. *Journal of Monetary Economics*, *30*. Retrieved from http://www.sciencedirect.com/science/article/pii/0304393292900649

Guiso, L., & Parigi, G. (1999). Investment and demand uncertainty. *Quarterly Journal of Economics*, *114*(1), 185–227. Retrieved from http://www.jstor.org/stable/2586951

Hamari, J., Sjöklint, M., & Ukkonen, A. (2013). The sharing economy: Why people participate in collaborative consumption. *Available at SSRN 2271971*. Retrieved from http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2271971

Hanemann, W. (1991). Willingness to pay and willingness to accept: how much can they differ? *The American Economic Review*, *81*(3), 635–647. Retrieved from http://www.jstor.org/stable/2006525

Hollard, G., Massoni, S., & Vergnaud, J. (2010). Subjective beliefs formation and elicitation rules: experimental evidence. Retrieved from https://halshs.archives-ouvertes.fr/halshs-00543828/

Horowitz, J. K. (2006). The Becker-DeGroot-Marschak mechanism is not necessarily incentive compatible, even for non-random goods. *Economics Letters*, *93*(1), 6–11. doi:10.1016/j.econlet.2006.03.033

Horowitz, J. K., & McConnell, K. E. (2002). A Review of WTA/WTP Studies. *Journal of Environmental Economics and Management*, *44*(3), 426–447. doi:10.1006/jeem.2001.1215

Hurd, M. D. (2009). Subjective Probabilities in Household Surveys. *Annual Review of Economics*, *1*, 543–562. doi:10.1146/annurev.economics.050708.142955

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. doi:10.1177/0956797611430953

John, N. a. (2013). The Social Logics of Sharing. *The Communication Review*, *16*(3), 113–131. doi:10.1080/10714421.2013.807119

Kaas, K., & Ruprecht, H. (2006). Are the Vickrey auction and the BDM-mechanism really incentive compatible? Empirical results and optimal bidding strategies in the case of uncertain willingness-to-. *… Results and Optimal Bidding Strategies in …*, (January), 37–56. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=878271

Kahneman, D., Knetsch, J., & Thaler, R. (1986). Fairness and the assumptions of economics. *Journal of Business*, *59*(4). Retrieved from http://www.jstor.org/stable/2352761

Kahneman, D., Knetsch, J., & Thaler, R. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy*, *98*(6), 1325–1348. Retrieved from http://www.jstor.org/stable/2937761

Kahneman, D., Knetsch, J., & Thaler, R. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *The Journal of Economic Perspectives*, *5*(1), 193–206. Retrieved from http://www.jstor.org/stable/1942711

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *49*(2), 1057–1072.

Karni, E., & Safra, Z. (1987). "Preference Reversal" and the Observability of Preferences by Experimental Methods. *Econometrica*, *55*(3), 675. doi:10.2307/1913606

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 37–41.

Kianifard, F., & Zelterman, D. (2000). Models for Discrete Data. *Technometrics*, *42*(3), 313. doi:10.2307/1271095

Knetsch, J. (1989). The endowment effect and evidence of nonreversible indifference curves. *The American Economic Review*, *79*(5), 1277–1284. Retrieved from http://www.jstor.org/stable/1831454

Knez, P., Smith, V., & Williams, A. (1985). Individual rationality, market rationality, and value estimation. *The American Economic Review*, *75*(2), 397–402. Retrieved from http://www.jstor.org/stable/1805632

Loomis, J. B. (2014). 2013 WAEA Keynote Address : Strategies for Overcoming Hypothetical Bias in Stated Preference Surveys. *Journal of Agricultural and Resource Economics*, *39*(1), 34–46.

Loughran, T. a., Paternoster, R., & Thomas, K. J. (2014). Incentivizing Responses to Self-report Questions in Perceptual Deterrence Studies: An Investigation of the Validity of Deterrence Theory Using Bayesian Truth Serum. *Journal of Quantitative Criminology*, *30*(4), 677–707. doi:10.1007/s10940-014-9219-4

Mann, H. ., & Whitney, D. . (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. Retrieved from http://www.jstor.org/stable/2238700

Manski, C. (2004). Measuring expectations. *Econometrica*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0262.2004.00537.x/abstract

Maselli, I., & Giuli, M. (2015). UBER: Innovation or déja vu?, (February), 1–3.

Nachar, N. (2008). The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution. *Tutorials in Quantitative Methods for Psychology*, *4*(1), 13–20. Retrieved from http://www.tqmp.org/doc/vol4-1/p13-20_Nachar.pdf

Perren, R., Administration, B., & Florida, C. (2015). *International Encyclopedia of the Social & Behavioral Sciences*. *International Encyclopedia of Social & Behavioral Sciences* (Second Edi., Vol. 4). Elsevier. doi:10.1016/B978-0-08-097086-8.64143-0

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science*, *306*(5695), 462–466. doi:10.1126/science.1102081

Rachel Botsman, R. R. (2010). Beyond Zipcar : *Harvard Business Review*, (October).

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk , Kolmogorov-Smirnov , Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, *2*(1), 21–33.

Rogers, B. (2015). The Social Costs of Uber, *1*.

Ross, L. E. E. (1977). The " False in Social Consensus Perception Effect ": An Egocentric Bias and Attribution Processes. *Journal of Experimental Social Psychology*. doi:doi:10.1016/0022-1031(77)90049-X

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. doi:10.3758/PBR.16.2.225

Rutström, E. E., & Wilcox, N. T. (2009). Stated beliefs versus inferred beliefs: A methodological inquiry and experimental test. *Games and Economic Behavior*, *67*(2), 616–632. doi:10.1016/j.geb.2009.04.001

Sayman, S., & Öncüler, A. (2005). Effects of study design characteristics on the WTA–WTP disparity: A meta analytical framework. *Journal of Economic Psychology*, *26*(2), 289–312. doi:10.1016/j.joep.2004.07.002

Schor, J. (2014). Debating the Sharing Economy, (October). Retrieved from http://www.geo.coop/sites/default/files/schor_debating_the_sharing_economy.pdf

Shogren, J., Shin, S., Hayes, D. J., & Kliebenstein, J. B. (1994). Resolving differences in willingness to pay and willingness to accept. *The American Economic …*, *84*(1), 255–270. Retrieved from

http://www.jstor.org/stable/2117981

Smith, V. L. (1982). Microeconomic System as an Experimental Science. *American Economic Review*, *72*(5), 923–955. Retrieved from http://www.jstor.org/stable/1812014

Sonnemans, J., Kuilen, G. Van De, & Wakker, P. P. (2009). Non-Bayesians : for Proper Scoring Rules, *76*(4), 1461–1489.

Thaler, R. H. (1985). Mental Accounting and Consumer Choice. *Marketing Science*, *4*(3), 199–214. doi:10.1287/mksc.4.3.199

Trautmann, S. T., & van de Kuilen, G. (2014). Belief Elicitation: A Horse Race among Truth Serums. *The Economic Journal*, n/a–n/a. doi:10.1111/ecoj.12160

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance*, *16*(1), 8–37. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.1982.tb03586.x/abstract

Weaver, R., & Prelec, D. (2012). Creating Truthtelling Incentives with the Bayesian Truth Serum. *Journal of Marketing Research*. doi:10.1509/jmr.09.0039

Willig, R. (1976). Consumer's surplus without apology. *The American Economic Review*, *66*(4), 589–597. Retrieved from http://www.jstor.org/stable/1806699

Zhao, J., & Kling, C. L. (2001). A new explanation for the WTA/WTPdisparity. *Economics Letters*, *73*(3), 293–300. doi:10.1016/S0165-1765(01)00511-0

Appendix

Appendix A. Survey Design

Starting the survey: respondents randomly assigned into 4 conditions: BTS-WTP, BTS-WTA, Control-WTP, Control-WTA.

BTS Condition

1) Instructions

Thank you for participating in this brief survey.

In this survey you will be asked to read a short description of a new product by UBER and answer **two questions**: your opinion and to estimate an average of other people with the same opinion.

For your complete answer you will have an opportunity to win **15 Euros** based on your answer's **"Truth Score".** Truth scoring, recently invented by an MIT professor and published in the academic journal Science, rewards you for truth telling. Even though it is only you who will know whether your answer is accurate, those telling the truth will score higher overall.

You are likely to win the prize if you answer truthfully by making sure you consider carefully the questions and answer honesty.

(Instructions are taken from Weaver & Prelec, 2012 article "Creating Truth-telling Incentives with the Bayesian Truth Serum". Details regarding the truth-scoring are avoided as Prelec (2004) suggests in his original Bayesian Truth Serum article.)

Short description

UberEATS is a delivery service offered by Uber from which you get delivered meals from local restaurants in 10 minutes or less. You can request a delivery, confirm it and track it.

Every week, there is a new menu, carefully organized from a selection of the best local restaurants to provide a variety of meal options for different tastes.

The meals are delivered by a messenger: by bike or by foot. You get fast messenger pickups and immediate deliveries.

Different meals at different prices are offered, plus a flat delivery fee regardless of the number of meals.

Additional information for the WTA condition:

As everybody can sign up as a bike messenger, imagine yourself becoming an UberEATS messenger. *Typically, the messenger gets around 80% of the delivery fee, the rest is for UBER.*

Q1 WTP: What is the maximum amount (the maximum delivery fee) in Euros that you are willing to pay to use to get a meal delivered?

[…..]

Q1 WTA: What is the minimum amount (the minimum delivery fee) in Euros that you are willing to accept to deliver meals?

[…..]

Q2: You indicated the maximum/minimum amount you are willing to pay/accept for delivery fee to be [insert price stated in Q1].

What do you feel is the price, in Euros, on average that other respondents would pay/accept?

[…..]

Control Group

1) **Instructions:**

Thank you for participating in this brief survey.

In this survey, you will be asked to read a short description of a new product by **UBER** and answer **a question.** A random winner will be elected with the possibility to win 15 Euros.

Please consider you response carefully and answer honestly.

Short description

UberEATS is a delivery service offered by Uber from which you get delivered meals from local restaurants in 10 minutes or less. You can request a delivery, confirm it and track it.

Every week, there is a new menu, carefully organized from a selection of the best local restaurants to provide a variety of meal options for different tastes.

The meals are delivered by a messenger: by bike or by foot. You get fast messenger pickups and immediate deliveries.

Different meals at different prices are offered, plus a flat delivery fee regardless of the number of meals.

Additional information for the WTA condition:

As everybody can sign up as a bike messenger, imagine yourself becoming an UberEATS messenger. *Typically, the messenger gets around 80% of the delivery fee, the rest is for UBER.*

Q1 WTP: What is the maximum amount (the maximum delivery fee) in Euros that you are willing to pay to use to get a meal delivered?

[…..]

Q1 WTA: What is the minimum amount (the minimum delivery fee) in Euros that you are willing to accept to deliver meals?

[…..]

Demographic questions for both conditions:

Q3

What is your gender?

- Male
- Female

Q4

What is your age?

[….] open-ended question

Q5

Where do you currently reside?

[….] open-ended question

Q6

What is the highest degree or level of school you have completed?

- High-school
- MO/HBO
- Bachelor
- Master
- Doctorate

Appendix B. R-code

```
btsData <- read.csv("bts.csv", header = TRUE, sep = ',')

controldata <-read.csv("control.csv", header = TRUE, sep = ',')

ttestBF(formula=BWTPWTA ~ GROUPBTS, data=btsData)

ttestBF(formula=CWTPWTA ~ GROUPC, data=controldata)

newbtsData <- btsData[which(btsData$BWTPWTA<15), ]
```

```
 ttestBF(formula=BWTPWTA ~ GROUPBTS, data=newbtsData)

newCData <- controldata[which(!(controldata$CWTPWTA==15 & controldata$GROUPC==1)),]

ttestBF(formula=CWTPWTA ~ GROUPC, data=newCData)
```