Cheap Talk with the Bayesian Truth Serum

Jae Joon Lee*

Stanford University

October 18, 2023

Abstract

Biased responses in survey studies could seriously harm and mislead our economic decision-making. To mitigate survey response bias, we suggest an alternative way of combining two existing strategies, cheap talk and the Bayesian Truth Serum (BTS). In our three proof-of-concept experiments, we found that our alternative approach, named the C-BTS, helps elicit more truthful survey responses even in situations where neither the BTS nor cheap talk alone works well enough, especially in the context of economic valuation of goods. By applying the C-BTS, we have also confirmed that AI-powered services are already significantly enhancing the well-being of our citizens.

Keywords: cheap talk, Bayesian Truth Serum, hypothetical bias, willingness to pay, willingness to accept
JEL: C83, C90, D90, D91, M31, O33

*The study was approved by Institutional Review Boards at Stanford University (eProtocol 68617). The author declares that he has no relevant or material financial interests that relate to the research described in this paper. The author thanks the financial support from the Stanford Digital Economy Lab. This RCT was registered as AEARCTR-0011560 (Lee, 2023). The views expressed herein are those of the author alone. This paper benefits greatly from the valuable comments of Erik Brynjolfsson, Drazen Prelec, Joshua Tasoff, and John Clithero. All errors are my own. Affiliation: The Stanford Digital Economy Lab at the Stanford Institute for Human-Centered Artificial Intelligence, Cordura Hall, 210 Panama Street, Stanford University, Stanford, CA 94305 (email: jaejoonl@stanford.edu)

1 Introduction

The data collected through surveys forms a groundwork for economic research and policymaking. Nevertheless, it has been widely discussed how reliable self-reported surveys are, and previous studies have found significant bias even in government-sponsored surveys (Moore et al., 1997; Bhandari et al., 2020; Davison et al., 2022; Comerford, 2023). There can be several potential sources of bias in survey data, such as the bias in sampling and data-collection mode (Kasprzyk, 2005; Bhandari et al., 2020). However, presumably, the most problematic bias might come from the survey responses themselves, because it is usually challenging to recover the unbiased responses using ex-post correction methods. Given how critically these survey data are used for economic research and policymaking, biased responses in surveys could seriously harm and mislead our economic decision-making.

There have been ongoing discussions on how to mitigate the bias in survey responses. Especially, in economic valuation, how to reduce so-called hypothetical bias, which can be defined as the difference between the stated value elicited from hypothetical choices in a survey and the revealed value from real choices (Murphy et al., 2005), has been extensively discussed in academia. Several approaches have been suggested on how to mitigate the bias, but presumably, the two most widely used approaches, especially in economic research, might be cheap talk (Cummings and Taylor, 1999) and the Bayesian Truth Serum (Prelec, 2004). Cheap talk is to let respondents themselves aware of their potential bias when answering hypothetical questions by directly providing the scripts on what hypothetical bias is and why it might occur, and then ask subjects to answer questions as if they were in a real situation (Cummings and Taylor, 1999). In contrast, the Bayesian Truth Serum (BTS) is a method of scoring the truthfulness of the responses and rewarding those who score higher with a bonus payment (Prelec, 2004). In this approach, truthfully stating one's belief is each respondent's best response to maximize the expected monetary payoff in the survey, given that everyone is responding truthfully. The efficacy of each approach has been extensively tested. While some studies have shown that each approach can mitigate the bias in survey responses successfully (Lusk, 2003; Carlsson et al., 2005; List et al., 2006; Weaver and Prelec, 2013; Frank et al., 2017), they were not effective enough in several other studies (Aadland and Caplan, 2006; Blumenschein et al., 2008; Barrage and Lee, 2010; Bennett et al., 2019; Menapace and Raffaelli, 2020).

We presume that efficacy varies across studies due to the structural limitations of each of the two approaches. Although cheap talk approach was named following a game theoretic situation where costless signaling (or truth-telling) can occur by having (at least) a partial alignment of interests between two parties, no mutually aligned interest exists in the context of a survey, as truth-telling is beneficial only to the researcher while it incurs only a mental cost for subjects who need to respond to the questions more carefully. Hence, there exists no incentive for subjects to answer survey questions sincerely, and this approach simply relies on the "good will" of the respondents. Due to this structural limitation, we expect cheap talk approach might not work well enough, especially when a survey question is more complex, and therefore, requires a higher mental cost, or when there exists a motive for providing deceptive responses for some reasons (e.g., self-image concern, etc.). In contrast, the BTS explicitly provides incentives for truth-telling in the form of monetary rewards, but as Prelec himself, who first proposed the BTS approach, pointed out, it may not eliminate all types of untruthfulness (Weaver and Prelec, 2013). Out of the three types of untruthfulness he mentioned, he confirmed the BTS can mitigate intentional deception and carelessness, but he noted that some sources of inauthenticity, which indicate the situation where answers are biased due to several reasons such as social norms or cognitive heuristics, may be fully unconscious and hard to be eliminated even by huge monetary rewards (Weaver and Prelec, 2013). Inauthenticity might not matter much for simple and obvious tasks. Indeed, in several studies where the BTS worked well, the context of the task was relatively simple (e.g., "From the list, how many items do you know?", etc.). However, we expect inauthenticity might matter much more when the context of the task gets complex, and many important surveys in economic reearch require respondents to consider multidimensional aspects of each task (e.g., contingent valuation on environmental goods).

To overcome the shortcomings of each approach described above, this study presents an alternative approach. We argue that cheap talk and the BTS can complement each other. If a surveyor provides pecuniary benefits using the BTS, respondents in cheap talk may have an incentive for answering questions more carefully as if they were in a real situation. Conversely, when the cheap talk script is included, respondents are more likely to be aware of their potential inauthenticity which could bias the responses seriously in the BTS. Hence, we suggest to combine cheap talk with the BTS for eliciting more truthful responses in surveys. We call this approach "cheap talk with the Bayesian Truth Serum (C-BTS)." We expect the C-BTS can elicit more truthful responses because it resembles a real-choice situation by having two core elements. First, in real choices, people think about their decision making in the context of their real life. Using the cheap talk part, the C-BTS can ask people to recall a real-life situation in making choices in a survey. Second, in real-life situations, a person's choice has a consequence that directly affects their utility. Using the BTS part, the respondent's choice can have a consequential effect on her utility in the form of the monetary payoff. In the implementation of the C-BTS in practice, we first provide a description of the BTS to respondents and ask them to answer several training questions so that they can better understand the BTS can actually reward their truthful responses. Then, we provide the cheap talk scripts on what hypothetical bias is and why it might occur. Finally, in each survey question, we remind respondents that the more accurate their answers are "as if they were in real situations", the more likely they are to receive an additional bonus payment in the survey.

To validate the C-BTS, we implemented three different types of experiments. First, as direct evidence of the efficacy of the C-BTS, we replicated the context of a previous study, Barrage and Lee 2010, in which neither cheap talk nor the BTS worked successfully in eliciting truthful responses. In this experiment, we asked respondents whether they would donate the $5 they earned during the experiment to children suffering from cancer or have it for themselves. We randomly assigned respondents into one of five experimental groups, including Real, Hypothetical, BTS, Cheap Talk, and C-BTS. Our results show that 61.3% of subjects chose to donate in the Hypothetical group, whereas only 29.7% of respondents chose to do so in the Real group. For the treatment groups intended for mitigating the hypothetical bias, 45.3% of subjects in the BTS group chose to donate and 47.9% of respondents in the Cheap Talk group chose to do so. In contrast, we confirmed that the choices in the C-BTS group were statistically indistinguishable from those in the REAL group; 29.5% of participants in the C-BTS group chose to donate the $5 instead of having it for themselves. This experiment demonstrates that the C-BTS can be successfully used even in situations where neither the BTS nor cheap talk alone work well enough.

In our second experiment, we tested the efficacy of the C-BTS in the measurement of the economic value of goods. More specifically, we tried to elicit the willingness-to-accept (WTA) for not using each of two popular social media apps, Facebook and Instagram for one week. For this purpose, we asked subjects to make a series of binary discrete choices. For instance, respondents were asked to choose either 'Yes' or 'No' to the question, "Would you be willing to avoid using Facebook for 1 week in exchange for getting $10?" Using 6 different dollar values, we derived a demand curve for each social media app by fitting a binary logit model to the respondents' responses. We found significant hypothetical bias again in this experiment. The estimated median WTA for not using Facebook for 1 week was $4.94 in the Hypothetical group, whereas it was $15.61 in the Real group. We also found that the median WTA in the BTS group ($5.03) was similar to that in the Hypothetical group. The median WTAs in the Cheap Talk group ($10.64) and in the C-BTS group ($11.58) were statistically different from the hypothetical responses, but they tended to understate the value of Facebook by around $4, compared to the Real group. We got consistent results for Instagram as well. There can be two hypotheses to explain the difference in responses between the Real group and the C-BTS group. First, it is possible that the C-BTS itself may not work well enough in this context. Second, the responses in the Real group can tend to overstate the true value of each social media app; To make choices consequential in the Real group, we informed subjects in this group that some of their choices could be fulfilled through the deactivation and monitoring process. If this process incurs significant costs for respondents, the responses in the Real group not only reflect the net value of each social media app but may also include deactivation and monitoring costs such as the loss of privacy. To test these hypotheses, we recruited another C-BTS group and asked subjects to hypothetically assume the same procedure given in the Real group. Then, the median WTA in this C-BTS group ($15.00) was statistically indistinguishable from the one in the Real group ($15.61). This result supports the hypothesis that the C-BTS itself works well and the previous difference between the C-BTS group and the REAL group came mainly from deactivation and monitoring costs. The deactivation and monitoring process has been a common practice for incentive compatibility in previous studies on measuring the value of social media apps (Corrigan et al., 2018; Mosquera et al., 2020; Brynjolfsson et al., 2019a; Allcott et al., 2020), but the costs from this process have been usually ignored. Our result shows that this practice can impose the risk of overstating the true value of social media apps. More generally speaking, our result shows that, while consequential choices have been treated as a gold standard in economic experiments, the elicited value could be biased even with consequential incentives if the procedure involves some significant costs to respondents.

Third, we implemented another experiment on economic valuation of 6 social media apps, using a different survey format. Conjoint analysis has been used extensively in business and marketing studies. In this experiment, we focus on a specific type of conjoint analysis, best-worst scaling (Wittenberg et al., 2016). Best-worst scaling (BWS) approach asks subjects to repeatedly select the best and worst options from several sets of alternatives (Flynn et al., 2007; Brynjolfsson et al., 2019b). For instance, in one BWS question, we asked respondents to choose the best and worst option among “Not using Facebook for the next 1 week”, “Not using Instagram for the next 1 week”, and “Earning $10 less for the next 1 week.” By fitting a statistical model such as a conditional multinomial logit to the subjects’ responses to the BWS questions, we can quantitatively measure relative utility from each of the 6 social media apps. The results showed that the C-BTS group was statistically indistinguishable from the Real group, while the BTS group and the Cheap Talk group were significantly different from the Real group, suggesting that the C-BTS can be successfully applied for economic valuation using the BWS format, a type of conjoint analysis. There have been debates on whether cheap talk scripts can truly mitigate bias, or they just introduce another bias unrelated to the hypothetical bias (Cummings and Taylor, 1999; List et al., 2006). In this experiment, by including several different dollar value items (e.g., earning \$10 less for the next 1 week), we could interpolate the implied willingness-to-pay (WTP) for each social media app in dollar terms. Using the implied WTPs, we checked the correlation between the size of hypothetical bias and the size of mitigated bias on 6 social media apps. We found the size of mitigated bias using the C-BTS was positively correlated with the size of hypothetical bias (ρ = 0.357), while there existed no correlation between cheap talk alone and hypothetical bias (ρ = -0.091). This result could serve as evidence that, unlike cheap talk alone, the C-BTS may help to mitigate the hypothetical bias, presumably by making subjects think more carefully, with financial incentives, about the real-choice situation.

The three proof-of-concept experiments discussed earlier show that, compared to either cheap talk or the BTS alone, the C-BTS can elicit more truthful responses in different contexts (donation and social media apps) and in different survey formats (binary discrete choices and BWS). Accordingly, as an application of the C-BTS, we implemented one more experiment to better understand one critical aspect of our current digital economy. It is widely accepted that artificial intelligence (AI) is reshaping how our economy works, but less understood how much AI is affecting our citizens' lives. One of the reasons could be that it is usually challenging to get reliable survey estimates in measuring the value of AI, especially using consequential choices. Therefore, this context may be especially suitable for the application of the C-BTS. In this experiment, we tried to measure the consumer value of 12 popular AI-powered services in daily life, such as real-time fraud alerts from one's credit card company, etc. As in our third experiment, we used BWS approach and interpolated the implied WTP for each AI-powered service. Although only 12 AI-powered services were included in this survey, the combined annual value of these services was approximately \$189.1 billion in the United States. This amounts to about 0.74% of U.S. GDP in 2022. As Brynjolfsson et al. 2019a pointed out, while the value of digital goods and technologies is not explicitly captured in traditional economic statistics, given that most of these AI-powered services are provided for free, this result suggests that AI is already significantly enhancing our citizens' quality of life.

This study makes contributions to the existing literature in two broad areas. First, we believe our suggested approach, the C-BTS, can contribute to the methodological advancement in mitigation strategies for the bias in survey responses, especially hypothetical bias in economic valuation. Although there is no consensus on the sources of hypothetical bias (Haghani et al., 2021; Lee and Hwang, 2016), several explanations have been suggested. Some studies have pointed out the lack of incentive compatibility might be a main driver of hypothetical bias (Haghani et al., 2021; Morkbak et al., 2014; Lewis et al., 2018; Buckell et al., 2020). Another group of studies have argued that moral or social desirability might prohibit respondents to answer survey questions more truthfully (Haghani et al., 2021; Andreoni, 1990; Leggett et al., 2003; Nunes and Schokkaert, 2003; Ding et al., 2005; Champ and Welsh, 2007; Olynk et al., 2010; Hainmueller et al., 2015; Smith et al., 2017; Svenningsen and Jacobsen, 2018; Menapace and Raffaelli, 2020; Sanjuan-Lopez and Resano-Ezcaray, 2020). The other studies also mentioned that cognitive biases, which naturally limits respondents to predict their actual behaviors, could be one of the most important factors causing hypothetical bias (Haghani et al., 2021; Loewenstein and Schkade, 1999; Frederick et al., 2002; Loewenstein et al., 2003). Several strategies have been suggested to mitigate hypothetical bias, such as time-to-think method (Haghani et al., 2021; Whittington et al., 1992), solemn oath (Jacquemet et al., 2017), honesty priming (de Magistris et al., 2013), cheap talk (Cummings and Taylor, 1999), and the Bayesian Truth Serum (Prelec, 2004). However, each strategy conceptually covers only a portion of the sources mentioned above, and thus previous studies have shown that its efficacy is context dependent. By combining two popular strategies, cheap talk and the Bayesian Truth Serum, we believe the C-BTS can more broadly address the main sources of hypothesis bias; By having the BTS component, the C-BTS can address the incentive compatibility problem. Our first experiment on donation decision making also demonstrated that it could help mitigate the bias from moral or social desirability concerns. In addition, by having the cheap talk component, the C-BTS can also mitigate cognitive biases arising from differences between hypothetical and real choice contexts. Consequently, we believe that the C-BTS can be applied to a wider variety of contexts than existing strategies in eliciting more truthful responses in surveys and mitigating hypothetical bias in economic valuation. Second, our study can contribute to the literature on better measurement of the digital economy as well. In the digital economy, greater penetration of internet access and rapid technological change is making many goods to be increasingly available for free, reflecting insignificant marginal costs of digital replication and distribution. As Brynjolfsson et al. 2019a pointed out, if some goods are consumed with a zero measured price, they can have zero measured value in the traditional economic statistics. Accordingly, there have been several attempts to better measure the consumer's valuation of free digital goods, such as social media apps, using choice experiments (Corrigan et al., 2018; Mosquera et al., 2020; Brynjolfsson et al., 2019a; Allcott et al., 2020). These studies introduced consequential incentives by monitoring respondents' social media app usage but usually ignored the cost associated with the process to respondents, such as the loss of privacy. Our second experiment demonstrated that this practice may risk overstating the true value of digital goods, which we believe will facilitate discussion on more robust ways of valuing digital goods. In addition, our application study on eliciting the consumer value of AI may shed light on how to better measure the impact of AI on our citizens' well-being. Despite the growing importance of artificial intelligence in our lives, there has been little research on how much value AI creates for consumers. A few studies (Zhang et al., 2022; Konig et al., 2022) have attempted to measure consumers' WTPs for certain aspects of AI. However, these studies have typically been limited in scope to specific designs of AI, rather than its general use, and have mainly relied on hypothetical responses. As we demonstrated, the application of the C-BTS could be helpful for future research in better measuring the impact of AI on our citizens' lives.

The rest of this paper is organized as follows: Section 2 walks through the conceptual framework. Section 3 describes a replication study of Barrage and Lee (2010). Section 4 validates the use of the C-BTS for measuring the willingness-to-accept (WTA) on two social media apps using binary discrete choices. Section 5 discusses the use of the C-BTS for measuring the willingness-to-pay (WTP) on six social media apps using best-worst scaling (BWS) approach. Section 6 demonstrates the application of the C-BTS for measuring the consumer value of AI-powered services in daily life. Finally, Section 7 concludes with implications, cautions, and directions for future research.

2 Conceptual Framework

As briefly discussed in Section 1, one of the challenges in eliciting more truthful responses in surveys is, some sources of hypothetical bias, such as cognitive biases (Haghani et al., 2021; Loewenstein and Schkade, 1999; Frederick et al., 2002; Loewenstein et al., 2003), could be partly or even fully unconscious to respondents. Indeed, several social psychologists have argued that beliefs activated in hypothetical situations are qualitatively different from beliefs in real contexts. (Ajzen and Sexton, 1999; Ajzen et al., 2004). Economists have usually emphasized the role of financial incentives, but if responses to hypothetical situations involve qualitatively different procedures than those used in real situations, it may not be helpful enough as respondents may not even be aware of their hypothetical bias (Weaver and Prelec, 2013). In this situation, social psychologists found that the ex-ante framing design using cheap talk scripts could be helpful by letting respondents to form beliefs and attitudes similar to those in a real choice situation (Ajzen et al., 2004). As already introduced, cheap talk is an approach that directly provide corrective entreaty on what hypothetical bias is and why it might occur, and then ask subjects to answer questions as if they were in a real situation (Cummings and Taylor, 1999; Ajzen et al., 2004). When respondents' beliefs are primed differently in hypothetical and real contexts (Wegener and Petty, 1995; Cummings and Taylor, 1999), several previous studies found that cheap talk scripts could mitigate hypothetical bias by inducing beliefs and attitudes aligned with those in a real choice situation (Cummings and Taylor, 1999; Ajzen et al., 2004; Jacquemet et al., 2011).

Nevertheless, cheap talk design cannot be a panacea to hypothetical bias (Jacquemet et al., 2011). One of the main reasons is, even if a respondent recognizes the difference between hypothetical and real contexts, it is entirely up to her free will whether she would respond to the survey questions carefully and sincerely, considering the real contexts. Therefore, it is expected that cheap talk alone may not be able to mitigate hypothetical bias especially if considering the real contexts carefully involves significant mental costs or if truthful responses can hurt her self-image due to social desirability concerns. In this situation, Weaver and Prelec 2013 demonstrated that the Bayesian Truth Serum (BTS) can eliminate the motives for providing careless or deceptive responses in surveys. As explained in Prelec 2004, the BTS consists of a scoring system that induces truthful answers from a sample of rational (i.e., Bayesian) expected value-maximizing respondents. It assigns high scores to answers that are more common than collectively predicted, with predictions drawn from the same population that generates the answers. Such responses are "surprisingly common," and the associated numerical index is called an information score. For instance, as given in our second experiment, we might ask: "Would you be willing to avoid using Facebook for 1 week in exchange for getting $10?" Each respondent provides a personal answer (Yes or No) and also a prediction of the empirical distribution of answers in the population (the fraction of people endorsing Yes or No). If 'Yes' endorsed by 50% of the population against a predicted frequency of 40%, then it is surprisingly common and the respondent who chose 'Yes' receives a high information score; if predictions averaged 60%, it would be a surprisingly uncommon answer, and hence a respondent receives a low score. For a better understanding of how the BTS works, let's assume the respondent's truthful answer is 'Yes.' Then, as a rational Bayesian expected value-maximizer, she should give higher estimates of the percentage of the population who prefer 'Yes' through Bayesian updating, because her own opinion is an informative "sample of one." However, she also knows that there are some people who prefer 'No' and they would give lower estimates of the proportion of the population who prefer 'Yes' in a similar way as she updates her belief. This causes the average of predictions in the population to be lower than her best guess for the true proportion of the population who prefer 'Yes.' In other words, from her perspective, the true popularity of 'Yes' is underestimated by the population. Hence, one's true opinion is also the opinion that has the best chance of being surprisingly common and giving a high score.

More formally, as given in Prelec 2004, if we denote answers and predictions by respondent r on a m multiple-choice question as $x^r = (x_1^r, \ldots, x_m^r)\,(x_k^r \in \{0,1\}, \Sigma_k x_k^r = 1)$ and $y^r = (y_1^r, \ldots, y_m^r)\,(y_k^r \geq 0, \Sigma_k y_k^r = 1)$, respectively, the population endorsement frequencies, $\bar{x_k}$, and the (geometric) average, $\bar{y_k}$, of predicted frequencies is given by:

$$\bar{x}_k = \lim_{n\to\infty} \frac{1}{n} \sum_{r=1}^{n} x_k^r,$$
$$\log \bar{y}_k = \lim_{n\to\infty} \frac{1}{n} \sum_{r=1}^{n} \log y_k^r$$

Then the information score for answer k is given by:

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

The total score for a respondent combines the information score with a separate score for the accuracy of predictions as follows:

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

This scoring rule leads to honesty being a Bayesian Nash equilibrium for $\alpha > 0$ and is a zero-sum game for $\alpha = 1$. Then, it is theoretically predicted that the BTS can elicit more truthful responses.

Nonetheless, an issue of inauthenticity can occur if a respondent's belief was already anchored to hypothetical contexts. Unlike the BTS, the C-BTS can help her align her beliefs with a real-life choice situation through corrective entreaty in cheap talk. If she believes that other respondents truthfully state their beliefs in the real context, then it is also her best response to provide truthful beliefs aligned with a real choice situation. In other words, truth-telling in the context of real-life situations by each respondent is a Bayesian Nash equilibrium. One potential pitfall might be that there could be an issue of multiple equilibria because the respondent can be aware of both hypothetical and real-life situations in the C-BTS. For instance, if a respondent believes that other respondents state their beliefs in the context of hypothetical situations, then it can be also her best response to do the same (i.e., each subject responding in the context of hypothetical situations can be an equilibrium). In fact, this issue of multiple equilibria can also matter in the original BTS as well as the C-BTS (Weaver and Prelec, 2013); For example, in the BTS, if a participant believes that all other subjects are responding in direct opposition to their true belief, then responding in the same way can be her best strategy as well. One way to solve the issue of potential multiple equilibria in the C-BTS could be to nudge respondents towards truth-telling in the context of real situations. For such a purpose, in the implementation of the C-BTS, we stated in each survey question that the more accurate their answers are as if they were in 'real situations', the more likely they were to receive an additional payment.

It is an empirical question whether such a nudge could work well and, more generally, whether the C-BTS can mitigate bias in survey responses more successfully than other strategies. From Section 3 to Section 5, we provide experimental evidence on this question.

3 Donation Decision Making

3.1 Background

One of the best ways to validate the C-BTS might be to replicate the experiment in a previous study where cheap talk and the BTS were individually ineffective and test the efficacy of the C-BTS in that context. For that purpose, we will focus on Barrage and Lee 2010 in this session. In their study, they implemented the experiments on donation decision-making in China. In one experiment, participants were asked whether they would choose to donate 30 Chinese yuan (about 4.5 U.S. dollars) from what they earned during the experiment to provide three tents for the China Foundation for Poverty Alleviation 's Disaster Relief division ("Tents"). In the other experiment, subjects were asked whether they would donate the same amount of money to staff the Pollution Victims Hotline for a day at the Center for Legal Assistance to Pollution Victims ("Hotline"). Their results show that neither cheap talk nor the BTS worked well in this context. Among the survey respondents in the "Tent" experiment, 48% chose to donate in the "real" treatment where they were actually required to donate. However, in cheap talk and the BTS treatment, 77% chose to donate, and this was not statistically different from the completely hypothetical responses (79%). In the "Hotline" experiment, compared to hypothetical responses (83%), both cheap talk (50%) and the BTS (55%) were somewhat helpful in mitigating bias, but there was still a gap between their responses and those in the real treatment (32%), as given in the appendix.

Donation decision-making has been widely used to test the efficacy of cheap talk and the BTS. Previous studies have usually shown that each strategy works well (Cummings and Taylor, 1999; List et al., 2006; Weaver and Prelec, 2013). Therefore, the results from Barrage and Lee 2010 raise the question of why neither worked well in this case. We hypothesize that the specific context may matter. The contexts in which either cheap talk or the BTS worked well, in previous studies, tended to be those related to the provision of public goods for the general public, such as support for maintenance of a system of pedestrian trails (Cummings and Taylor, 1999), environmental protection (List et al., 2006), and art projects (Weaver and Prelec, 2013). In contrast, in Barrage and Lee 2010, the donation was targeted for a specific group with emergent needs, such as those requiring tents for disaster relief. This context may evoke more empathy in participants than those used in previous studies, and if they choose not to donate, there may be a higher chance that it could harm their self-images, such as perceiving themselves as immoral.

In practice, it was hard for us to replicate a previous study (Barrage and Lee, 2010) exactly as it was implemented in another country. Instead, based on the hypothesis discussed above, we chose a situation that could evoke more empathy in participants to replicate the context of their study. More specifically, in our experiment, we asked respondents whether they would donate to support 'children suffering from cancer.' The detailed experimental procedure and design are as follows.

3.2 Design and Procedure

The experiment was conducted on Connect, CloudResearch's recently launched survey platform. CloudResearch has focused on providing toolkits to ensure superior data quality in surveys, so we expected their new survey platform might work well for obtaining more reliable responses[1]. We recruited 388 participants and randomly assigned them to one of five treatment groups, including (a) Real, (b) Hypothetical, (c) BTS, (d) Cheap Talk, and (e) C-BTS group, in March 2023. Each respondent received a participation fee of $1 after finishing

[1] Indeed, in our pilot, we confirmed that the attention check failure rate was significantly lower than that of some other existing platforms.

the experiment.

(a) In the Real group, we first asked subjects to implement real-effort tasks[2] to mitigate the potential house money effect. Once subjects completed the real-effort tasks, they were asked to answer a main survey question. We began with emphasizing this was for a real stake and asked each subject to choose one of two options: keeping all the money they just earned or donating $5 from what they just earned to St. Jude Children's Research Hospital to help children suffering from cancer. We provided additional background that since its opening in 1962, St. Jude Children's Research Hospital has increased the overall childhood cancer survival rate from 20% to over 80%. Following Barrage and Lee 2010, we explained to the subjects that the decision would be made by a majority voting process.

(b) Unlike the Real group, subjects in the Hypothetical group were not asked to implement the real-effort tasks. Instead, we asked subjects to suppose that they had earned an additional $6 by completing an extra survey in this experiment with hard work and care. After emphasizing that the question was hypothetical, we asked each subject to choose one of the same two options given for the Real group.

(c) For the BTS group, we started by describing the BTS algorithm to the respondents. We informed that this algorithm would give participants a higher score the more accurately they answer questions after careful thinking, and we would use this score to rank the survey respondents and award a bonus of $20 to the top 5% of responders. Then we asked them to answer 10 random training questions out of 20[3]. After answering the main question on donation decision-making, we also asked each participant what percentage of people would have chosen to donate in order to calculate the collective prediction for the BTS score.

(d) For the Cheap Talk group, we first presented the cheap talk scripts to subjects. We attempted to generally follow the scripts used for Cummings and Taylor 1999 and List et al. 2006 but made some modifications for certain issues. First, in these studies, the scripts stated the direction and size of hypothetical bias observed in another study. For instance, in Cummings and Taylor 1999, it was mentioned that in a recent study, 38% of

[2] We asked subjects to review a survey question the author was working on for another study. This survey question consisted of five sentences. For each of the three most confusing sentences, we asked participants to provide (i) reasons why it could be confusing and (ii) suggestions on how to improve it. Consequently, each subject was required to answer 6 questions in total. For each complete answer, they can earn $1, resulting in the earning of up to $6 from real-effort tasks.

[3] To gather preliminary data on how people responded to the training questions, we recruited another pool of 300 participants on Connect, three days before implementing the main experiment. Each respondent was asked to answer 10 random ones from a pool of 20 questions. Therefore, we obtained 150 responses on average for each of the 20 training questions, and we used this data to pre-calculate the BTS score for each choice. During the main experiment, we showed this BTS score to the subjects for their choices.

respondents voted to donate in a hypothetical referendum, while 25% of them voted to do so in a real situation. Similarly, in List et al. 2006, it was stated that people overstated their actual willingness-to-pay by 150 percent in the hypothetical auction in a previous study. We worried that specifying the direction and size of hypothetical bias might cause another bias by experimenter demand effects. Therefore, in our cheap talk scripts, without specifying the direction and size of hypothetical bias, we simply mentioned, "What we observed was a clear difference in the responses across the groups on average." Second, in their cheap talk scripts, sometimes strong expressions such as "quite a difference" were used, but we attempted to use more neutral expressions to avoid causing another bias (e.g., saying simply "different" instead of "very different"). The detailed cheap-talk scripts used in this experiment is provided in the appendix. After presenting the cheap talk scripts to subjects, we asked subjects to answer the same question given to other groups. We emphasized that the question was hypothetical, but asked them to please answer as if they were in real situations.

(e) Finally, for the C-BTS group, we first provide the description of the BTS algorithm with training questions given in (c), and then, we provided the cheap talk scripts given in (d). Afterwards, the subjects were asked to provide answers to the same question given to other groups. We stated, "This question is hypothetical, but please answer as if you were in real situations. You are more likely to get a real bonus payment of $20 if you answer the question accurately as if you were in real situations." Then, we additionally asked a question for the collective prediction as already explained above.

In each group, there was an attention-check question[4] after answering the main question. 8 out of 388 respondents failed to pass the attention check. After dropping those who failed the attention check, we had 380 responses for the analyses (Real:74, Hypothetical:80, Cheap Talk:73, BTS:75, and C-BTS:78).

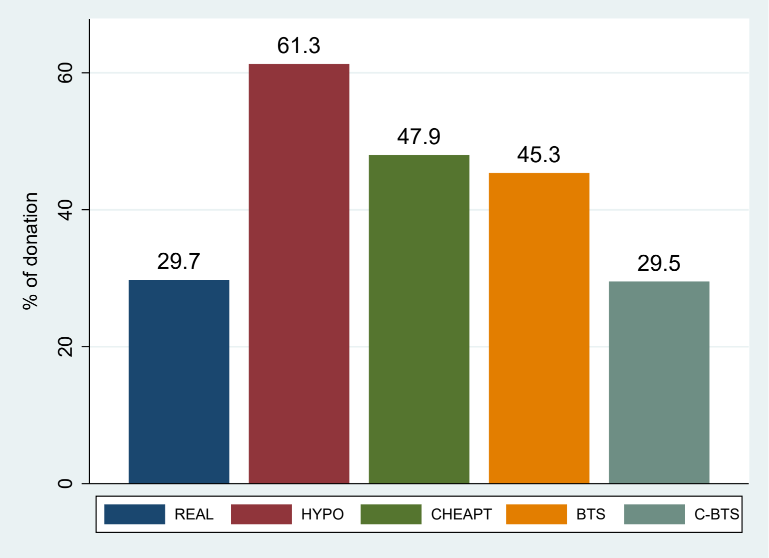
3.3 Results and Discussion

The result of this experiment is summarized in Figure 1. First, consistent with previous studies, we found significant hypothetical bias in donation decision-making. Specifically, 61.3% of subjects chose to donate in the Hypothetical group, whereas only 29.7% of respondents chose to do so in the Real group, implying that more than twice as many respondents were willing to donate in the hypothetical context compared to the real context. Second, it seems

[4] In this attention check, we asked subjects to choose the correct statement about the question they just answered among "I was asked whether I would like to donate $5 that I earned in this experiment", "I was asked whether I would like to donate $6 that I earned in this experiment", and "The contribution will be used for the purpose of helping old adults with cancer."

that cheap talk and the BTS, respectively, helped to mitigate the hypothetical bias to some extent in this experiment. 45.3% of subjects in the BTS group chose to donate and 47.9% of respondents in the Cheap Talk group chose to do so. However, there was still a significant gap between the responses in these groups and those in the real group. Using the two-sample test of proportions, the responses in the Real group were significantly different from those in either the Cheap Talk group or the BTS group (p-value: 0.0234 and 0.0493 for cheap talk and the BTS, respectively). These results are consistent with the "Hotline" experiment in Barrage and Lee 2010. In contrast, the responses in the C-BTS group, 29.5%, were very close to those in the Real group, 29.7%, and the difference was not statistically significant (p-value: 0.9739). Since some participants provided feedback that the attention check question was too difficult, for robustness check, we additionally analyzed the responses of all participants without using the attention-check question, but we could not observe statistically meaningful difference, as given in the appendix.

Figure 1: The Experimental Results on Donation Decision-making



In this experiment, we tried to replicate the context of a previous study (Barrage and Lee, 2010), in which neither cheap talk nor the BTS worked well. Our results show that, even in such a context, the C-BTS can elicit more truthful responses as good as those in real-choice situations. We presume that the success of the C-BTS in this experiment was due to the complementarity between cheap talk and the BTS as discussed earlier. People usually tend to perceive themselves as good persons who are willing to help someone in need, rather than someone who prioritizes small financial gain. In this situation, cheap talk can help individuals realize what they would actually do in a real situation, but choosing not to donate in this experiment can be like admitting themselves as selfish persons, so there is no reason to answer in that way without any incentive. In contrast, the BTS cannot fully eliminate inauthenticity, so respondents might not have even realized what they would do in a real situation, because they believe they might choose something benevolent as good persons. By combining cheap talk with the BTS, the C-BTS might have helped respondents to consider the real context and truthfully respond due to the financial incentive which can counteract social desirability concern.

4 Measuring the WTA Using Binary Discrete Choices

4.1 Background

Survey-based approaches have been widely used for economic valuation, especially on non-market goods for which market prices do not exist. For instance, contingent valuation has been a popular practice to measure the value of environmental goods or the impact of externalities. In this experiment, we are trying to measure the value of another type of goods for which a price does not exist in the market. Due to insignificant marginal costs of digital replication and distribution, digital goods are increasingly available for free. If some goods are consumed with a zero price, we cannot usually measure the value of those goods in the market. Consequently, there have been several attempts to apply the survey-based approach for the measurement of the value of digital goods (Corrigan et al., 2018; Mosquera et al., 2020; Brynjolfsson et al., 2019a; Allcott et al., 2020). In addition, the use of digital goods involves multidimensional aspects of life, such as communicating with friends, having access to news feeds, spending leisure time, etc., which could cause several existing strategies for mitigating hypothetical bias to not function properly. Therefore, we aim to verify the efficacy of the C-BTS in measuring the value of digital goods in this experiment. Specifically, we will focus on measuring the willingness-to-accept (WTA) for not using each of the two most popular social media apps, Facebook, and Instagram, for one week. There are several different survey formats available to elicit the WTA for goods, but we will use the binary discrete choice format in this experiment, as Brynjolfsson et al. 2019a did in their study, for two main reasons. First, simply choosing one of the two options is one of the simplest question formats, and therefore there is relatively less chance of bias in the valuation as it is less likely to cause confusion among participants[5]. Second, since the question we used in our first experiment was also a binary discrete choice, we would like to demonstrate how to extend this approach to economic valuation.

4.2 Design and Procedure

In this experiment, as a screening question, we first asked each participant if they had used Facebook (or Instagram)[6] in the past month. If participants answered that they have used the social media app in the screening question, they were asked to choose between 'Yes' or 'No' to a question such as "Would you be willing to avoid using Facebook for 1 week in exchange for $10?" We asked each subject to make a series of six binary discrete choices, each time using a different dollar value ($1, $4, $7, $10, $20, or $50) in a random order. We recruited 742 respondents and randomly assigned them to one of five treatment groups, including (a) Real, (b) Hypothetical, (c) BTS, (d) Cheap Talk, and (e)C-BTS group, in March 2023. Each respondent received a participation fee of $1.2 after finishing the experiment.

(a) First, in the Real group, at the beginning of the experiment, we informed subjects that we would randomly pick 1 out of every 100 respondents and one of her choices could be fulfilled. For instance, if a respondent chose to avoid using Facebook for 1 week in exchange for getting $10, she was asked to provide her Facebook page URL, and to deactivate her Facebook account for one week. We informed that the experimenter would keep checking her Facebook page and pay her $10 if it kept remaining inactive for 1 week. On the other hand, if she did not make such a choice, she could keep using Facebook, but she could not get the bonus payment of $10. For other treatment groups, we have not mentioned such a procedure for making choices consequential, and the remaining experimental procedures were similar to the ones in the first experiment. (b) In the instruction for the Hypothetical group, we mentioned that the questions were hypothetical, but asked the subjects to answer every question accurately. (c) In the BTS group, just like the first experiment, we provided a description of the BTS with training questions. Then, in each question, we emphasized that the more accurate their answers were, the more likely they were to receive a bonus payment of $20. After answering six main questions, each subject was asked to predict what percentage of people would have chosen 'Yes' to one random main question she just answered in order to calculate the collective prediction for the BTS score. (d) In the Cheap Talk group, we first

[5] In fact, in our pilots, we found that this format provided more reasonable estimates in measuring the value of goods than other formats such as open-ended questions like the Becker-DeGroot-Marschak method (BDM), or the multiple price list (MPL), especially in online experiments.

[6] Each social media app was presented in a random order.

presented cheap talk scripts which were similar to the ones used in the first experiment but customized for the valuation of social media apps, as given in the appendix. After presenting the cheap talk scripts to subjects, in each question, we asked subjects to answer the question as if they were in real situations. (e) In the C-BTS group, after presenting a description of the BTS with training questions and cheap talk scripts as described above, we emphasized that the more accurate their answers were as if they were in real situations, the more likely they were to receive a bonus payment of $20 in this experiment. Participants were also asked to answer one random question for the collective prediction as given in the BTS group.

We used 2 types of attention checks in this experiment. First, there was a specific question[7] that was directly intended for the attention check. Second, we dropped subjects who provided inconsistent answers[8]. After screening questions on the previous usage of each social media app and attention checks, we had 594 respondents on Facebook (Real:126, Hypothetical:124, Cheap Talk:113, BTS:122, and C-BTS:109) and 505 respondents on Instagram (Real:111, Hypothetical:113, Cheap Talk:94, BTS:98, and C-BTS:89).

4.3 Results and Discussion

For the estimation of the demand for each social media app, we assume a random utility model with a logistically distributed error term. This allows us to express the observed choices within a binary logit model as given below, where x indicates the amount of money offered in each choice and y indicates a binary choice of whether to keep using each social media app for 1 week in exchange for getting x (y=1) or not (y=0).

$$\mathrm{P}(y) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

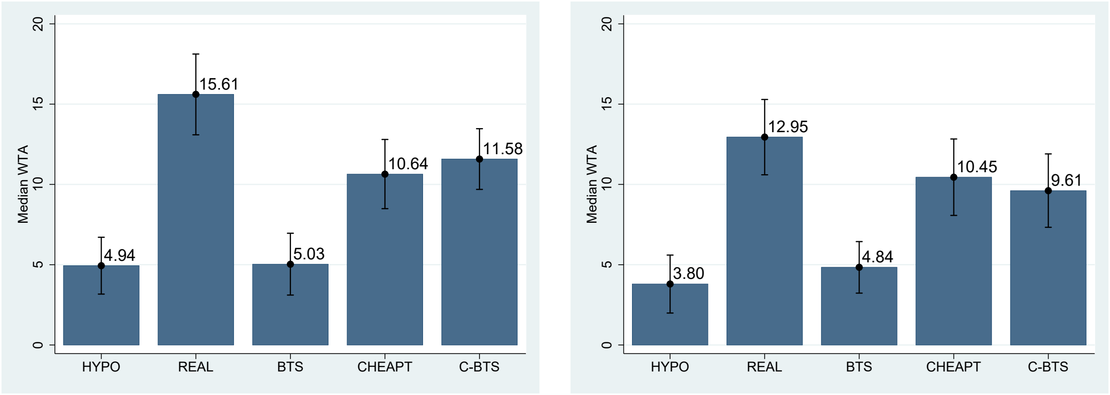
The parameters can be estimated using closed form maximum likelihood procedures. Then, using the estimated parameters, we derived the demand curve for each social media app from each of 5 treatment groups, as given in the appendix.

For comparisons across treatment groups, we focus on the median willingness-to-accept

[7] In this question, we asked subjects to choose the correct statement about the question they just answered among “I was asked whether I would be willing to avoid using Facebook (or Instagram) for 1 day in exchange for getting some amount of money”, “I was asked whether I would be willing to avoid using Facebook (or Instagram) for 1 week in exchange for getting some amount of money”, and “I was asked whether I would be willing to avoid using Facebook (or Instagram) for 1 month in exchange for getting some amount of money.”

[8] For instance, when a subject was asked whether she would be willing to avoid using Facebook for 1 week in exchange for getting some amount of money, responding ‘No’ for $1 and $4, ‘Yes’ for $7, and again ‘No’ for $10, $20, and $50 did not make sense, and we considered such responses as attention check failures.

Figure 2: The Median WTA for Each Social Media App



(a) Facebook

(b) Instagram

(WTA) in each group as a representative summary statistic, since it is not influenced by certain outliers. In the binary logit model given above, the median willingness-to-accept (WTA) for not using each social media app for 1 week is given by the value of x^* that makes $P(y) = 0.5$ or $\alpha+\beta x^* = 0$ which leads to $x^* = \text{-}\alpha/\beta$. After estimating the median WTA in each group, we calculated its 95% confidence intervals using bootstrapping. The results are given in Figure 2.

First, we observed a significant hypothetical bias for both apps. While the median WTAs for Facebook and Instagram were only about \$4.94 and \$3.80, respectively in the Hypothetical group, they were \$15.61 and \$12.95 in the Real group. The BTS does not seem helpful in mitigating the hypothetical bias, with the median WTAs for Facebook and Instagram being \$5.03 and \$4.84, respectively. Using the permutation test, we found the responses in the BTS group were statistically indistinguishable from those in the Hypothetical group (p-value: 0.957 for Facebook and 0.380 for Instagram). The median WTAs for Facebook and Instagram were \$10.64 and \$10.45 in the Cheap Talk group, showing statistical difference from those in the Hypothetical group (p-value: 0.000 for both apps). However, they were still different from the median WTAs in the Real group (p-value: 0.004 for Facebook and 0.132 for Instagram). The results from the C-BTS group were similar to those from the Cheap Talk group. The median WTAs for Facebook and Instagram were \$11.58 and \$9.61 in the C-BTS group, with statistical difference from those in the Hypothetical group (p-value: 0.000 for both apps). Nevertheless, we still observed some gap in median WTAs between the Real group and the C-BTS group (p-value: 0.013 for Facebook and 0.039 for Instagram).

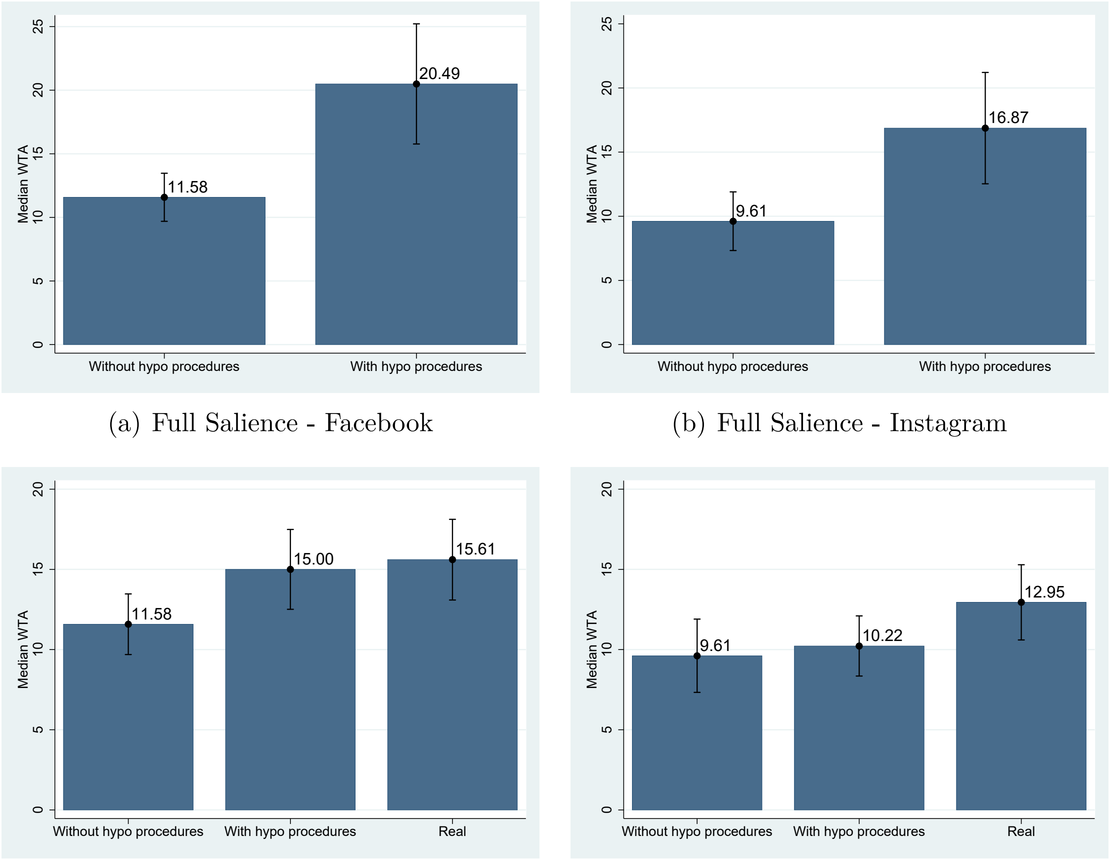
These results raised a question about what caused such a gap between the Real and the C-BTS group. We established two hypotheses: First, similarly to how the efficacy of other mitigating strategies is context-dependent, the C-BTS may not work well enough in this specific context. Second, while the C-BTS can successfully mitigate hypothetical bias, the responses in the Real group may have overstated the true value of each social media app. As explained earlier, to make choices consequential in the Real group, we informed the respondents that the experimenter would collect their Facebook (or Instagram) page URL and monitor whether they had truly deactivated each social media app by checking their social media pages regularly for one week. If participants were reluctant to provide their social media URL or to bother to deactivate their account, or did not want to be monitored by the experimenter, the responses in the Real group may not only reflect the net value of each social media app but also include deactivation and monitoring costs, such as the loss of privacy. Indeed, in the post-experiment feedbacks, many participants expressed a strong reluctance to provide their personal information and to be monitored by the experimenter[9].

To test these hypotheses, we conducted additional pilots. Our strategy was to ask another group of respondents for the C-BTS treatment to assume a hypothetical deactivation and monitoring procedure which was exactly the same as the one given to the Real group. To investigate the potential impact of deactivation and monitoring costs, such as privacy concerns, we increased the salience of the hypothetical deactivation and monitoring procedure in one pilot. Unlike in the Real group where the procedure was explained only once in the instructions, we repeatedly explained the procedure in every question to increase its salience in this pilot. After screening and attention checks, we had 67 and 57 respondents for Facebook and Instagram, respectively. The results are given in Figure 3 (a) and (b). We observed a surprisingly huge increase in the WTA for each app. Compared to the previous results, the median WTA for Facebook increased from \$11.58 to \$20.49 in the C-BTS group. Similarly, the median WTA for Instagram increased from \$9.61 to \$16.87 in the C-BTS group. These results provide evidence that the cost of the deactivation and monitoring process for consequentiality could be potentially huge enough.

We cannot directly compare the WTAs from this pilot with those in the Real group in the main experiment, as we made the cost for deactivation and monitoring process more salient in this pilot than in the main experiment. Accordingly, we implemented another pilot. In this pilot, we explained the hypothetical deactivation and monitoring procedure only once in the instructions, as we did for the Real group in the main experiment. After screening and

[9] For instance, one participant mentioned, “The only reason I would not take your offer at some level is because in real life I would not be willing to let you monitor my Facebook or Instagram accounts or give you access to my personal data for any amount mentioned.”

Figure 3: The Change in the Median WTAs with Hypothetical Procedures



(a) Full Salience - Facebook

(b) Full Salience - Instagram

(c) Partial Salience - Facebook

(d) Partial Salience - Instagram

attention checks, we had 107 and 92 respondents for Facebook and Instagram, respectively. The results are given in Figure 3 (c) and (d).

For Facebook, the median WTA increased from \$11.58 to \$15.00 with the hypothetical procedure treatment, which resulted in a statistically significant change in the responses (p-value: 0.033). In contrast, this median WTA was statistically indistinguishable from the median WTA in the Real group, \$15.61 (p-value: 0.727). These results support our second hypothesis. Although the C-BTS can successfully elicit truthful responses as effectively as consequential incentives, the true value of Facebook in the Real group was overstated because of the costs associated with a deactivation and monitoring process, such as privacy concerns. According to these results, we estimate that such a process caused the WTA for Facebook to be overstated by approximately \$3.5 in the Real group.

On the other hand, for Instagram, the WTA obtained in this pilot was not statistically different from that obtained in the main experiment where there was no description of the hypothetical procedure. This result is questionable because we have already confirmed that the costs of the deactivation and monitoring processes could potentially be significant enough for Instagram as well. We suspect that these results may have resulted from our inability to induce the same level of attention to the hypothetical procedure for Instagram as in the Real group in the main experiment. Indeed, our instruction was mainly written in terms of Facebook, with every sentence and example mentioning Facebook first and Instagram only mentioned in parentheses, as given in the appendix. Studies in psychology and neuroscience show that people's thought processes differ in real and hypothetical situations, and that different parts of the brain are activated in each case (Camerer and Mobbs, 2017). In real situations, people tend to think more actively and pay more attention, whereas in hypothetical situations, people tend to be more passive in their thinking. Accordingly, when respondents were asked to hypothetically assume a situation in our pilot, it was likely that they might have paid less attention to the word given in parentheses compared to real choices. Then, a hypothetical situation that respondents have imagined could be mainly the collection of personal information on Facebook, with less focus on Instagram. To test this hypothesis in another pilot, we wrote the instructions in the opposite way, with Instagram mentioned first in every sentence and example, and Facebook mentioned only in parentheses. The results support our hypothesis. With this modified instruction, while the median WTA for Instagram was still approximately $3.5 lower than that for Facebook in the Real group, which is consistent with previous results, the median WTA for Instagram was only about $0.5 lower than that for Facebook in the C-BTS group[10]. These results suggest that the specific wording used in the instruction clearly affects the valuation of each social media app in the C-BTS group more than in the Real group, and our previous pilot results on Instagram may have been mainly caused by our failure to induce the same level of attention to the hypothetical procedure for Instagram as in the Real group, not because the costs for deactivation and monitoring procedure did not matter.

In short, the results of our experiment support the hypothesis that the C-BTS can successfully elicit more truthful responses in economic valuation. Our pilot results also demon-

[10] Due to the budget constraint, we recruited only 35 and 32 respondents for the Real and the C-BTS groups in this pilot. In the Real group, the media WTAs were $11.95 and $8.47 for Facebook and Instagram, respectively. In contrast, they were $22.97 and $22.48 in the C-BTS group. Due to small sample sizes, we cannot assume these absolute dollar values are statistically valid enough. However, comparing Facebook and Instagram in each group can still be valid because most people use both social media apps, and each respondent answered for both Facebook and Instagram in this pilot. In other words, comparing Facebook and Instagram in each group can be considered a quasi-within-subject design, which requires a much smaller sample size than a between-subject design.

strated that the difference in the median WTAs between the C-BTS group and the REAL group was mainly due to deactivation and monitoring costs associated with consequential incentives. As mentioned earlier, there have been several recent attempts to measure the value of social media apps using the survey-based approach, and the deactivation and monitoring process has been a common practice to make choices consequential (Corrigan et al., 2018; Mosquera et al., 2020; Brynjolfsson et al., 2019a; Allcott et al., 2020). Nevertheless, the costs of deactivation and monitoring borne by respondents have generally been ignored in the measurement of the WTAs for social media apps. Our results suggest that such a practice can cause potentially significant bias in the measurement of the value of social media apps. More generally, we believe our results can serve as a warning that even when consequential choices are involved in economic experiments, the elicited value could be biased if the procedures for consequentiality impose significant costs on respondents.

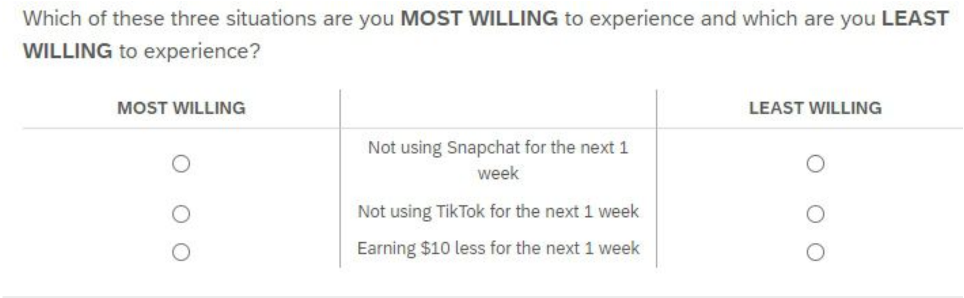
5 Measuring the WTP Using the Best-Worst Scaling

5.1 Background

In the previous section, we tested the efficacy of the C-BTS in eliciting the WTAs for social media apps. However, literature has pointed out that there exists a huge disparity between the WTA and the willingness-to-pay (WTP), and some studies have found that the WTA and the WTP might involve different cognitive processes and brain activities (De-Martino et al., 2009; Chapman et al., 2017). Accordingly, in our third proof-of-concept experiment, we tried to validate the C-BTS in eliciting the WTPs for social media apps. In addition, although the binary choice format we used in the second experiment was simpler to implement and less likely to cause confusion among participants, it may not be the most popular survey format for economic valuation in practice. To verify the potential broader applicability of the C-BTS, we decided to validate its efficacy in one of the most popular survey formats for economic valuation. Conjoint analysis has been used extensively in business and marketing studies in practice. In this experiment, we applied a specific type of conjoint analysis, namely, best-worst scaling (Wittenberg et al., 2016). Best-worst scaling (BWS) asks consumers to repeatedly select the best and worst options from several sets of alternatives (Flynn et al., 2007; Brynjolfsson et al., 2019b), as shown in Figure 4. As Brynjolfsson et al. 2019b pointed out, collecting more information, both within the choice set and across sequential choice sets, for each respondent could make this approach more efficient compared to the binary-choice approach, which elicits only one decision. Therefore, we tested the validity of using the

C-BTS in the BWS format to measure the WTPs for social media apps.

Figure 4: A Sample BWS Question Used in the Experiment



Which of these three situations are you **MOST WILLING** to experience and which are you **LEAST WILLING** to experience?

MOST WILLING		LEAST WILLING
○	Not using Snapchat for the next 1 week	○
○	Not using TikTok for the next 1 week	○
○	Earning $10 less for the next 1 week	○

5.2 Design and Procedure

In this experiment, we measured the willingness-to-pay (WTP) for six popular social media apps including Facebook, Instagram, Pinterest, Snapchat, Twitter, and TikTok [11]. In addition, we considered seven monetary values including $1, $4, $7, $10, $15, $20, and $50. As a result, there were 13 items in this experiment, including 6 social media apps and 7 monetary values. Since we aimed to measure the WTP, the specific wording used for each app was "Not using a social media app (e.g., Facebook) for the next 1 week," and for each monetary value it was "Earning a specific amount of money (e.g., $10) less for the next 1 week." We included 3 items in each BWS question as illustrated in Figure 4. Using a balanced incomplete block design (BIBD), we created 26 BWS questions for this experiment [12]. After a screening question on the previous usage of each social media app, participants answered up to 10 random BWS questions. The three items in each question were also presented in a random order. Again, we had 5 treatment groups, including Real, Hypothetical, BTS, Cheap Talk, and C-BTS group. Each subject received a participation fee of $2.4 for completion.

In the Real group, we first asked participants to perform the same real-effort tasks used in the first experiment[13]. Then, we notified that we would randomly pick 1 out of every

[11] While other social media apps such as YouTube may be more popular than some of these, we chose these apps because they allowed us a deactivation and monitoring process for consequentiality in the Real group.

[12] Of those, two questions contained only monetary items, which made the choices trivial. To improve data collection efficiency, we excluded those two questions from the survey and later imputed the responses based on the assumption that the subjects had answered them properly.

[13] Each subject was required to answer 6 questions in total. For each completed answer, they could earn experimental currency of $10, resulting in earning up to experimental currency of $60 from real-effort tasks.

100 respondents and would exchange the experimental currency they just earned for real money. To make choices consequential, we adopted an incentive compatible conjoint ranking mechanism (Lusk et al., 2008). We instructed respondents that, out of the three items in one random question they answered, we would randomly choose one item to fulfill. The item they were most willing to experience would be selected with a 67% (2/3) chance, while the item they were least willing to experience would never be selected. The item they were neither most willing nor least willing to experience would be selected with the remaining 33% (1/3) chance. Then, for instance, in a sample question given in Figure 4, if the option "Not using Snapchat for the next 1 week" was chosen, subjects were asked to deactivate their Snapchat account for one week, but no money was deducted. In contrast, if the option "earning $10 less for the next 1 week" was chosen, $10 would be deducted from what they earned from the real-effort tasks, but they could continue to use Snapchat and TikTok freely. For other treatment groups, we have not mentioned such a procedure for making choices consequential, and the remaining experimental procedures were similar to the ones in two previous experiments. For the Cheap Talk group and the C-BTS group, we used almost identical cheap talk scripts to those used in the second experiment because both experiments were focused on measuring the value of social media apps[14]. After a screening question on the previous usage of each social media app and attention checks[15], we had 668 respondents for analyses (Real:158[16], Hypothetical:137, Cheap Talk:121, BTS:130, and C-BTS:122).

5.3 Results and Discussion

In the analyses of the preference for each social media app included in this experiment, we fit the conditional logit model to the responses from each treatment group. One assumption

[14] We conducted this experiment first, followed by the two other experiments described earlier. We attempted to use more neutral expressions in the scripts to avoid causing any bias, but in this experiment, the word "very" was mistakenly included in one sentence: "Respondents' assessments of how much they dislike losing access to certain goods in a 'hypothetical' setting were very different from their assessments in 'real' situations." We implemented several pilots but could not find evidence that the inclusion of this single word significantly affected the responses. However, to make things sure, for the other experiments described earlier, we dropped the word "very".

[15] In this experiment, 14 questions had two out of three options related to monetary amounts. We used these questions for attention checks. For instance, it would not make sense if a respondent chose "earning $15 less for the next 1 week" as the option she would be most willing to, or if she chose "earning $1 less for the next 1 week" as the option she would be least willing to because it is clearly contradictory that she preferred "earning $15 less" to "earning $1 less."

[16] In our preliminary analyses, the responses in the Real group tended to be noisier than those in other groups, so we additionally recruited subjects for the Real group so that the sample size for this group could be slightly larger than other groups.

underlying the estimation is how we assume respondents make the best and worst choices among the items given in each question. There are three standard models, including maxdiff, marginal, and marginal sequential (Aizaki, 2021; Flynn et al., 2008; Hensher et al., 2015; Louviere et al., 2015). In our analysis, we use the maxdiff model, which has been shown to be useful in demonstrating the properties of some estimators in the BWS (Marley and Pihlens, 2012). The maxdiff model assumes that people choose the best and worst items to induce the greatest utility difference among all possible pairs. Therefore, to estimate preferences on social media apps, we generate all possible pairs of best and worst choices for each question and fit the conditional logit model to the responses on which pair was actually chosen as the best and worst items. More specifically, the systemic component of the utility for our analyses can be written as the equation given below[17], and we estimated $P(chosen) = \Lambda(v)$.

$$
\begin{aligned}
v =& \beta_1 Facebook + \beta_2 Instagram + \beta_3 Pinterest + \beta_4 Snapchat + \beta_5 Twitter + \beta_6 TikTok \\
&+ \beta_7 dollar4 + \beta_8 dollar7 + \beta_9 dollar10 + \beta_{10} dollar15 + \beta_{11} dollar20 + \beta_{12} dollar50
\end{aligned}
$$

The results are given in Table 1. The estimated coefficient can be interpreted as a negative value of the relative utility from each item, as each item used in our experiment was "losing access to" a specific social media app or "earning less by" a specific amount of money. For instance, the results from the Real group are shown in column (2). When we set the relative utility of \$1 as 0, the utility from having \$4 is 0.544, and the relative utilities from Facebook and Instagram are 0.340 and 0.282, respectively. At a glance, looking at the coefficients given in Table 1, it seems that we have obtained results approximately consistent with those in the second experiment; There exists a significant difference in coefficients between the Hypothetical group given in column (1) and the Real group given in column (2). While the coefficients from the BTS group given in column (3) tend to be similar to those from the Hypothetical group, the coefficients from (4) the Cheap Talk group and (5) the C-BTS group tend to be closer to those from the Real group. Using the Chow test, we statistically examined the structural differences in coefficients across groups. Indeed, the Hypothetical group responses were structurally different from the Real group responses (p-value: 0.000). For each mitigating strategy, the BTS was not statistically distinguishable from the Hypothetical group (p-value: 0.5434). For cheap talk, while it tends to be statistically different

[17] In estimation, the best item is coded as 1, the worst item is coded as -1, and the items not shown or not chosen are coded as 0. "Earning \$1 less for the next 1 week" was used as a baseline, so it was dropped from the equation.

from the Hypothetical group (p-value: 0.0061), it was also different from the Real group (p-value: 0.0071). In contrast, the C-BTS group was structurally indistinguishable from the Real group (p-value: 0.5308), while it was clearly different from the Hypothetical group (p-value: 0.0002). These results suggest that the C-BTS is superior to the two other mitigating strategies and more valid for eliciting more truthful WTPs in the BWS format.

The relative utility from each item given in Table 1 is a little bit hard to interpret, so we decided to interpolate the dollar value of each item. We assumed the quadratic utility functional form[18] and fitted it to the estimated relative utility from each of the seven dollar values given in Table 1. This quadratic functional form fits very well to our data (adjusted R-squared > 0.98 for all of the 5 treatment groups), but we observed, for a few items, the interpolated dollar values could be negative, which is economically nonsensical since respondents would not have used such social media apps in such cases. A more reasonable assumption might be that these goods provide a very small non-negative amount of utility close to zero. Accordingly, we assumed the quadratic utility functional form bounded below by zero to interpolate the implied dollar values. The results are given in Table 2.

There are several findings using the interpolated dollar values given in Table 2. First, we observed a huge difference between the WTAs and the WTPs. For instance, in the Real group, while the median WTAs for Facebook and Instagram were \$15.61 and \$12.95, respectively, from our second experiment, the estimated WTPs for Facebook and Instagram were only about \$2.93 and \$2.67. The result that the WTAs are generally higher than the WTPs is consistent with the results from previous literature, but the difference between them appears strikingly huge in our experiments[19]. Second, as we already observed from Table 1, the interpolated dollar values on social media apps in the BTS group look very close to the Hypothetical group. Consistent with our second experiment, it seems the BTS alone is not sufficient enough to mitigate the bias in survey responses in measuring the value of social media apps. Third, whereas the Cheap Talk group was structurally different from the Real group in the conditional logit model estimation given in Table 1, the converted dollar values on several social media apps in the Cheap Talk group tend to be reasonably close to those in the Real group. Fourth, we got similar results from the C-BTS group as well; The converted dollar values of several social media apps, including Facebook, Instagram, and TikTok, tend

[18] A quadratic functional form satisfies basic properties of preferences, such as monotonicity and convexity, and it is widely assumed in portfolio theory. We chose this functional form because it fits our data better than some other possible functional forms, such as a logarithmic utility function.

[19] We presume that the specific wording used for dollar value items, such as "Earning less," made it difficult for participants to accept such a situation, resulting in low values being placed on social media apps as measured in our study.

Table 1: The Estimated Relative Utility from Each Item

	(1)	(2)	(3)	(4)	(5)
Items	HYPO	REAL	BTS	CHEAP	C-BTS
Facebook	-0.0294	-0.340	0.199	-0.371	-0.431
	(0.115)	(0.109)	(0.118)	(0.125)	(0.123)
Instagram	0.138	-0.282	0.309	-0.268	-0.336
	(0.129)	(0.119)	(0.129)	(0.136)	(0.136)
Pinterest	0.716	0.0821	0.982	0.260	0.445
	(0.165)	(0.155)	(0.174)	(0.166)	(0.211)
Snapchat	0.488	-0.481	0.925	0.268	-0.187
	(0.193)	(0.185)	(0.184)	(0.223)	(0.183)
Twitter	0.0900	-0.446	0.147	-0.677	-0.152
	(0.136)	(0.118)	(0.130)	(0.134)	(0.130)
TikTok	0.0572	-0.311	0.228	-0.213	-0.476
	(0.145)	(0.141)	(0.142)	(0.153)	(0.159)
Losing $4	-0.794	-0.544	-0.582	-0.753	-0.555
	(0.118)	(0.105)	(0.117)	(0.123)	(0.118)
Losing $7	-1.798	-1.300	-1.549	-1.741	-1.314
	(0.147)	(0.118)	(0.144)	(0.147)	(0.133)
Losing $10	-2.374	-1.515	-2.123	-2.087	-1.626
	(0.172)	(0.122)	(0.167)	(0.156)	(0.141)
Losing $15	-3.795	-2.542	-3.689	-3.500	-2.604
	(0.217)	(0.143)	(0.222)	(0.205)	(0.163)
Losing $20	-5.806	-3.584	-5.468	-5.259	-3.827
	(0.320)	(0.173)	(0.305)	(0.290)	(0.207)
Losing $50	-8.605	-4.845	-7.329	-7.141	-5.172
	(0.651)	(0.265)	(0.456)	(0.462)	(0.318)
Losing $1	-	-	-	-	-
Observations	8,178	8,448	8,142	7,122	6,954

* The standard errors of the estimated coefficients are given in parentheses.

Table 2: Interpolated Dollar Values

Items	HYPO	REAL	BTS	CHEAP	C-BTS
Facebook	1.73	2.93	1.44	2.74	3.32
Instagram	1.26	2.67	1.13	2.42	2.91
Pinterest	0.00	1.08	0.00	0.82	0.00
Snapchat	0.30	3.57	0.00	0.80	2.29
Twitter	1.40	3.41	1.58	3.70	2.14
TikTok	1.49	2.80	1.36	2.25	3.51
Losing $1	1.65	1.44	2.00	1.60	1.51
Losing $4	3.93	3.86	3.69	3.95	3.85
Losing $7	6.98	7.48	6.67	7.24	7.29

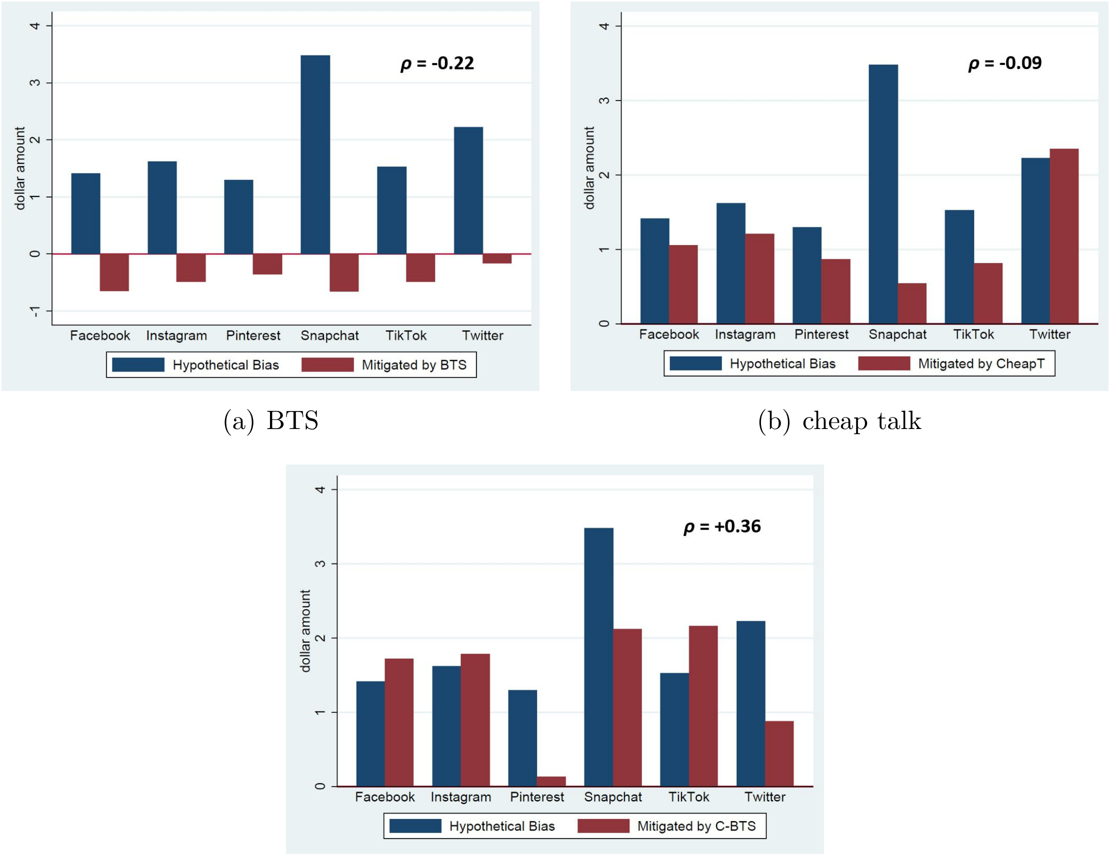
to provide reasonable estimates of the WTPs, but for some other social media apps, like Pinterest, Snapchat, and Twitter, there exist a gap of more than $1 in the estimated WTPs between the C-BTS and Real groups. We presume that the observed discrepancy could be mainly attributed to noise in the model estimation, rather than the C-BTS not working well. Indeed, the coefficients of Snapchat and Twitter in the C-BTS group and the coefficient of Pinterest in the Real group were not statistically significant in Table 1. Finally, unlike in our second experiment, it seems that the cost of the deactivation and monitoring procedure, such as privacy concern, did not seriously bias the responses in this experiment, as there is no clear pattern of overstatement of the WTPs in the Real group compared to the Cheap Talk or C-BTS group. We presume there can be at least two possible reasons; First, the survey format might matter. It might have been less clear how participants' responses would affect their privacy in the BWS format. For instance, if participants chose "stop using a social media app" as their best option, there was a 67% (2/3) chance, not a 100% chance, they would be asked to undergo the deactivation and monitoring process. Even in cases where they did not choose "stop using a social media app" as their best or worst option, they still had a 33% (1/3) chance of being asked to do it. In addition, out of the 24 questions in this experiment, 10 questions had at least two items on social media use out of three options. Comparing two social media use items within a single question could reduce concerns about privacy, as participants could be asked to undergo the deactivation and monitoring process regardless of which option they choose. Second, as discussed earlier, the elicited WTPs in this experiment were much smaller than the WTAs from the second experiment. For example, the total WTP for Facebook in the Real group was less than $3. Even if privacy concerns affected some participants' responses, the effect size may be minimal.

In short, we found that cheap-talk-related structures, including cheap talk alone and the C-BTS, tended to help elicit more accurate responses that were closer to the responses in the Real group, while the BTS alone did not perform well. Nonetheless, there have been some discussions on whether cheap-talk-related structures could actually mitigate hypothetical bias or if they just introduce another independent bias, which accidentally offsets the hypothetical bias. The results from previous studies are mixed; Some studies, such as Cummings and Taylor 1999, found that cheap talk scripts could truly mitigate hypothetical bias, while other studies, such as List et al. 2006, have concluded that they might introduce another bias which counteracts hypothetical bias. To test these hypotheses, one could analyze the correlation between the size of hypothetical bias and the amount of reduced bias resulting from each mitigating strategy. If a mitigating strategy can effectively reduce hypothetical bias, then we would expect to observe a positive correlation between them. In Table 2, we interpolated the approximate dollar value of each social media app in each group. Accordingly, we can calculate the size of hypothetical bias, which can be defined as the difference in the dollar value of each item between the Real group and the Hypothetical group. One issue could be that the size of the bias from interpolation might be different between groups. Consequently, we calculated the adjusted size of hypothetical bias using a difference-in-differences style approach. For instance, the adjusted size of hypothetical bias on Facebook is given by (the interpolated dollar value of Facebook in the Real group - the interpolated dollar value of Facebook in the Hypothetical group) - (the interpolated dollar value of $1 in the Real group - the interpolated dollar value of $1 in the Hypothetical group)[20]. We calculated the adjusted size of hypothetical bias on each social media app using such an approach. We could also calculate the amount of reduced bias resulting from each mitigating strategy in such a way. Then we investigated the correlation between the size of hypothetical bias and the amount of reduced bias resulting from each bias-mitigating strategy used in our experiment. The results are summarized in Figure 5.

As discussed earlier, it seems clear that the BTS alone is not helpful enough to reduce hypothetical bias in our specific context. The correlation coefficient between the size of hypothetical bias and the amount of reduced bias using the BTS was negative, -0.22. While cheap talk induces reasonable approximation of the WTP for some social media apps in terms of the interpolated dollar values, we found almost no correlation (ρ = -0.09) between the size of hypothetical bias and the amount of reduced bias using cheap talk. In contrast, using the C-

[20] (\$2.93 - \$1.73) - (\$1.44 - \$1.65) = \$1.41

Figure 5: The Correlation between Hypothetical Bias and Reduced Bias by Each Strategy



(a) BTS

(b) cheap talk

(c) C-BTS

BTS, we found the positive correlation between the size of hypothetical bias and the amount of reduced bias ($\rho = +0.36$). In this experiment, with only six social media apps included, it may be premature to draw definitive conclusions about the correlation between the size of hypothetical bias and the amount of reduced bias resulting from each strategy. Nevertheless, we interpret these results as showing the potential that, unlike other bias-mitigating strategies, the C-BTS can effectively reduce hypothetical bias by inducing participants to carefully consider real-choice situations through the use of monetary incentives.

Overall, the results from this experiment show that the C-BTS can provide reasonable estimates of the true WTPs using the BWS format; The estimated results from the conditional logit model indicate that, unlike other treatment groups, the responses from the C-BTS group were not statistically distinguishable from those from the Real group. In addition, using the interpolated dollar values, we found the positive correlation between the size of hypothetical bias and the amount of reduced bias resulting from the C-BTS. Given that the BWS is a type of conjoint analysis, which is a popular format for economic valuation, these results demonstrate the potential for the C-BTS to be more widely applicable in practice to elicit more truthful responses in survey studies.

6 Measuring the Consumer Value of AI-powered Services

6.1 Background

In economics, survey-based approaches have been primarily used for evaluating the value of non-market goods. However, as discussed earlier, with the emergence of the digital economy, digital goods without market prices are increasingly affecting people's lives, leading to a growing need for their economic evaluation. Therefore, in recent studies, survey-based approaches have been widely adopted for economic valuation of social media apps (Corrigan et al., 2018; Mosquera et al., 2020; Brynjolfsson et al., 2019a; Allcott et al., 2020). Following these studies, we also used social media apps to validate the efficacy of the C-BTS for economic valuation in Sections 4 and 5. In this section, we will demonstrate one application of the C-BTS to better understand another critical aspect of our current digital economy. Artificial intelligence (AI) is already impacting various aspects of our lives, and it is predicted to become an essential part of our lives in the near future. Nevertheless, there have been few studies on the economic impact of AI, and many of those studies have mainly focused on the productivity aspect or have been qualitative analyses rather than quantitative research on its impact on our citizens' well-being. There have been a few previous studies measuring consumers' willingness-to-pay for specific aspects of AI (Zhang et al., 2022; Konig et al., 2022). However, these studies are typically based on hypothetical survey responses, which have limitations. These limitations are presumed to arise from the difficulty of introducing consequential incentives for certain AI features in choice experiments. Since we have identified the potential for the C-BTS to be used as a substitute for consequential incentives in survey studies in previous sections, we have decided to apply the C-BTS to measure the consumer value of AI-powered services in daily life.

6.2 Design and Procedure

Nowadays, AI has become ubiquitous in our lives, making it difficult to measure the consumer value of AI using an exhaustive list of all AI-powered services. Instead, we have selected the 12 most commonly used AI-powered services in our daily lives, as given in Table 3.

Table 3: 12 AI-powered Sevices Considered in This Study

AI-powered services
1.Real-time fraud alerts from a credit card company or bank
2.Real-time traffic information on a mobile map (Google or Apple Map, etc.)
3.Email spam filters
4.Predictive search terms on search engines (Google, Bing, Yahoo, etc.)
5.Friend recommendations on social media (Facebook, Instagram, etc.)
6.Personalized ads on social media (Facebook, Instagram, etc.)
7.Voice assistants (Siri, Alexa, Google Assistant, etc.)
8.Face ID or fingerprint scans to unlock a smartphone
9.Autofocus feature to take a photo using a smartphone
10.Content recommendation on video or music streaming apps
11.Instant chatbots for immediate customer service
12.Real-time matching on ridesharing apps (Uber, Lyft, etc.)

We used the BWS survey format, which was employed in our third experiment, as it enables us to collect data more efficiently, when compared to the binary-choice format. In this experiment, our focus was on examining the annual WTP for each AI-powered service. Thus, the specific wording used for each service item was: "Not using each AI-powered service for the next 1 year." Additionally, we considered 9 monetary values, including \$1, \$5, \$10, \$20, \$50, \$100, \$500, \$1,000, and \$5,000. The wording employed for these monetary values was: "Earning some amount of money (e.g., \$100) less for the next 1 year." We created 70 questions employing a balanced incomplete block design (BIBD)[21]. The experimental procedure was the same as the one conducted for the C-BTS group in our previous experiments; We first provide the description of the BTS algorithm with training questions, and then, we provided customized cheap talk scripts for evaluating the value of AI-powered services, as outlined in the appendix. Then, participants were asked to answer up to 15 random questions after responding to a screening question regarding their previous usage of each AI-powered service.

[21] In the implementation, we excluded five questions that solely involved monetary-value items for more efficient data collection. Subsequently, we imputed the responses for these questions based on the assumption that the subjects had answered them correctly.

For each question, we emphasized: “The more accurately you answer, as if you were in real situations, the more likely you are to receive an additional payment of $20 in this experiment.” As our goal is to measure the consumer value of AI-powered services in the population, we attempted to recruit a representative sample of U.S. adults who are over 18 years old in this experiment. We recruited 266 respondents using an automatic census-matched template provided by the Connect platform, which matched participants based on U.S. citizenship, gender, age group, ethnicity, race, income, and education level. After dropping subjects who failed in attention checks[22], we had 216 respondents for analyses.

6.3 Results and Discussion

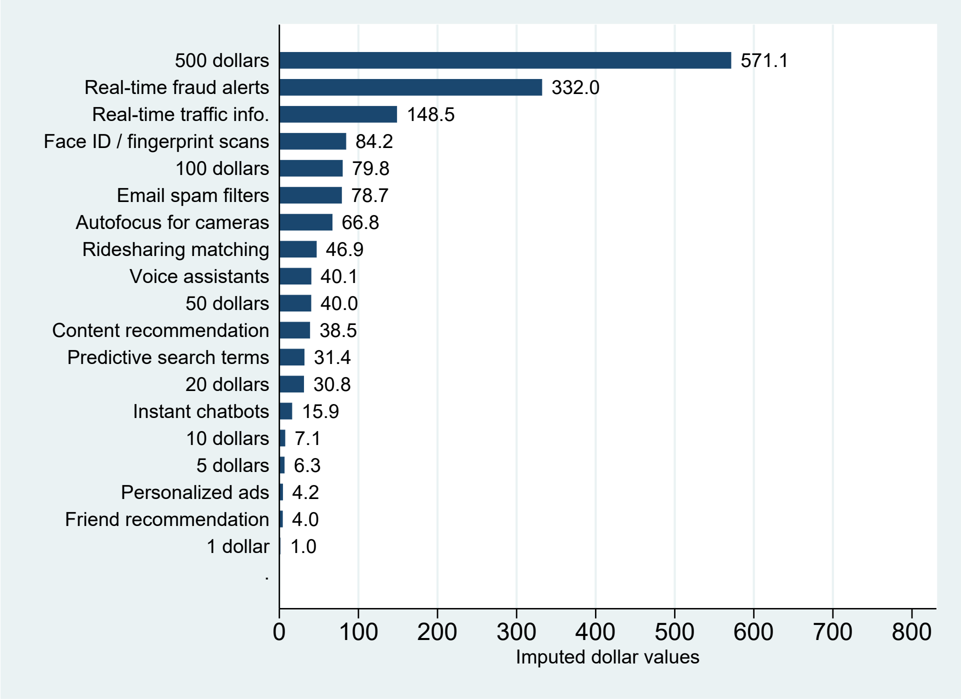
We fit the conditional logit model to the responses, and the results are given in the appendix. However, the estimated coefficients, which can be interpreted as the relative utility from each AI-powered service, are a little bit hard to interpret, so we interpolated[23] the dollar value of each item as we did in Section 5. The results are given in Figure 6.

We found that several AI-powered services are highly valued, even in terms of the WTP, which tends to be significantly lower than the WTA. For instance, the interpolated dollar value of “getting real-time fraud alerts from one’s credit card company or bank” is estimated to be around $332.0 per year. In our screening question, 80.9% of respondents whose information matched the census data reported using this service in the past year. As a back-of-the-envelope calculation, with a total adult population of 258.3 million in the U.S., the value of this AI-powered service alone amounts to approximately $69.4 billion per year. Similarly, the value of “using real-time traffic information on one’s mobile maps” is about $32.8 billion per year, and the value of “using any email spam filters” is about $19.0 billion per year. Considering these services are usually provided for free, these estimates highlight the significant contribution of AI-powered services to the well-being of citizens, despite the fact that the benefits from these digital goods may not be explicitly captured in economic statistics (Brynjolfsson et al., 2019b). This phenomenon is not limited to “free” digital goods. Even in physical products with a positive market price, like smartphones, AI plays a significant role in greatly enhancing our lives. For example, the “autofocus” feature, valued at $66.8 per year, was first introduced in the iPhone 3GS in 2009. However, the price of the

[22] Among the total of 65 questions in this experiment, 21 questions had 2 out of 3 options related to monetary amounts, and we used these questions as attention checks, just as we did in our previous experiment.

[23] Unlike the previous experiment in Section 5, where we used a quadratic utility function for interpolation, in this experiment, we assumed a log utility function because it better fits our data (R-squared: 0.9876). When fitting a log utility function, we dropped $5,000 as it resulted in greater bias.

Figure 6: The Interpolated WTPs for AI-powered Services



iPhone decreased by $100[24] compared to the previous model, the iPhone 3. Similarly, the "Face ID" feature, which was first introduced in the iPhone X in 2017, is valued at $84.2 per year. Despite a price increase of $300[25] compared to the previous model, the iPhone 8, mainly due to significantly improved physical features (such as a much larger screen, changes in a form factor, etc.), considering that people use smartphones for multiple years, this single feature itself can almost offset the price increase of the product. We found that some AI-powered services such as instant chatbots, personalized ads, and friend recommendations have relatively lower value. However, even though we evaluated the value of only 12 out of countless AI-powered services, they still provide an annual value of approximately $189.1 billion to the adult population in the United States. This amount corresponds to about 0.74% of the total U.S. GDP (approximately $25.46 trillion in 2022). Nonetheless, even this figure may represent only a small portion of the overall value that AI brings to society.

We additionally investigated demographic heterogeneity in the WTPs for AI-powered services, by gender, age, income, and education level[26]. We found that females generally

[24] For the 16GB model, the price of the iPhone 3GS was $199, while the price of the iPhone 3 was $299.

[25] For the 64GB model, the price of the iPhone X was $999, while the price of the iPhone 8 was $699.

[26] Due to the limited sample size, we performed binary classification based on the median value for each

tend to place higher value on AI-powered services included in this experiment compared to males. Older people tend to value certain services, such as "real-time fraud alerts " and "email spam filters," significantly more than younger people. Among the low-income group, there was a relatively higher valuation of "Face ID " and "content recommendation," whereas the high-income group tended to value "autofocus for cameras" relatively more. Similarly, among the low-education group, there was a relatively higher valuation of "Face ID" and "content recommendation," whereas the high-education group tended to value "email spam filters" and "ridesharing" relatively more. The detailed results are provided in the appendix.

7 Conclusion

Survey-based research plays a important role in collecting valuable data, which are then used for making critical economic and business decisions. Therefore, if there exist biases in the survey responses, it could seriously harm and mislead our decision-making. Several bias-mitigating approaches have been suggested, but each approach conceptually covers only a portion of the potential sources of hypothetical bias. This study suggests an alternative approach, the C-BTS. Due to the complementary relationship between cheap talk and the BTS, the C-BTS can more comprehensively address the main sources of hypothesis bias, such as the lack of incentives, social desirability concerns, and cognitive biases. In our proof-of-concept experiments, we show that the C-BTS can elicit more truthful responses even in situations where neither the BTS nor cheap talk alone work well enough. We have also confirmed that the C-BTS works well in different formats (binary discrete choices and BWS format) and with different contexts (donation decision-making and economic valuation of digital goods). We presume that the success of the C-BTS in our experiments can be attributed to its structure being similar to real choices: The cheap talk component induces people to think about their decision-making in the context of their real life, and the BTS component ensures that their choices have consequences that directly affect their utility.

We would like to provide some advice on the application of the C-BTS in practice. Firstly, we recommend avoiding the use of specific numbers (e.g., people overstated their actual WTP by "200 percent" in a hypothetical situation) or some strong expressions (e.g., there was a "huge" discrepancy between real and hypothetical responses in previous studies) in cheap

demographic factor as follows: (1) gender - male and female, (2) age - old (equal to or more than 38 years old) and young (less than 38 years old), (3) income: high (personal income: equal to or greater than $40,000) and low (personal income: lower than $40,000), and (4) education level: high (at least associate degree) and low (some college but no degree or without any college attendance) talk scripts, as they can introduce biases. In the C-BTS, if bias occurs from the cheap talk component, we are concerned that it may be even further amplified by the monetary incentives provided in the BTS component. Second, we believe that the process of participants recalling real-life situations through cheap talk scripts requires a more complex cognitive process, compared to simply recognizing the fact that the BTS can actually reward truthful responses. Therefore, we recommend showing the BTS instruction first and presenting cheap talk scripts just before participants respond to the survey questions, so that they can better recall real-life situations during the survey process. Indeed, in our pilots, we found that the C-BTS did not perform well enough when we reversed the order of the BTS and cheap talk instructions. Finally, there can be concerns about whether the C-BTS should always be used instead of cheap talk or the BTS in every survey. We believe that the answer to this question may vary depending on the context of each survey. On one hand, if the survey content is very simple and pertains to a clear situation , and therefore it is certain that there will be no potential cognitive bias arising from the difference between real and hypothetical situations, using the BTS alone would be sufficient. On the other hand, if it is guaranteed that survey participants will respond to the questions as carefully and truthfully as possible, without any carelessness or intentional deception, using cheap talk alone would be fine. However, in general, we recommend using the C-BTS for surveys where the context is multidimensional and there is a possibility of bias arising from several unknown sources.

The C-BTS successfully elicited more truthful survey responses in this study, but additional robustness checks will be necessary. In future research, comparing the efficacy of the C-BTS with other approaches that were not tested in this study (e.g., time-to-think method, honesty priming, solemn oath, etc.) would be helpful. Additionally, validating the use of the C-BTS in a wider variety of contexts (such as health, environment, political opinions, etc.) would also be beneficial. We hope that this study contributes to the ongoing discussions on better ways to mitigate biases in survey studies.

References

Aadland, David and Arthur J. Caplan, "Cheap talk reconsidered: New evidence from CVM," *Journal of Economic Behavior Organization*, 2006, *60* (4), 562–578.

Aizaki, H., "Support.BWS: Tools for Case 1 Best-Worst Scaling," 2021. https://CRAN.R-project.org/package=support.BWS.

Ajzen, Icek and James Sexton, "Depth of processing, belief congruence, and attitude-behavior correspondence," *Dual-process theories in social psychology*, 1999, pp. 117–138.

_ , Thomas C. Brown, and Franklin Carvajal, "Explaining the Discrepancy between Intentions and Actions: The Case of Hypothetical Bias in Contingent Valuation," *Personality and Social Psychology Bulletin*, 2004, *30* (9), 1108–1121. PMID: 15359015.

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow, "The Welfare Effects of Social Media," *American Economic Review*, March 2020, *110* (3), 629–76.

Andreoni, James, "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving," *The Economic Journal*, 06 1990, *100* (401), 464–477.

Barrage, Lint and Min Sok Lee, "A penny for your thoughts: Inducing truth-telling in stated preference elicitation," *Economics Letters*, 2010, *106* (2), 140–142.

Bennett, Richard, Kelvin Balcombe, Philip Jones, and Andrew Butterworth, "The Benefits of Farm Animal Welfare Legislation: The Case of the EU Broiler Directive and Truthful Reporting," *Journal of Agricultural Economics*, 2019, *70* (1), 135–152.

Bhandari, Anmol, Serdar Birinci, Ellen R. McGrattan, and Kurt See, "What Do Survey Data Tell Us about US Businesses?," *American Economic Review: Insights*, December 2020, *2* (4), 443–58.

Blumenschein, Karen, Glenn C. Blomquist, Magnus Johannesson, Nancy Horn, and Patricia Freeman, "Eliciting Willingness to Pay Without Bias: Evidence from a Field Experiment*," *The Economic Journal*, 2008, *118* (525), 114–137.

Brynjolfsson, Erik, Avinash Collis, and Felix Eggers, "Using massive online choice experiments to measure changes in well-being," *Proceedings of the National Academy of Sciences*, 2019, *116* (15), 7250–7255.

_ , _ , W. Erwin Diewert, Felix Eggers, and Kevin J Fox, "GDP-B: Accounting for the Value of New and Free Goods in the Digital Economy," Working Paper 25695, National Bureau of Economic Research March 2019.

Buckell, John, Justin S. White, and Ce Shang, "Can incentive-compatibility reduce hypothetical bias in smokers' experimental choice behavior? A randomized discrete choice experiment," *Journal of Choice Modelling*, 2020, *37*, 100255.

Camerer, C and D. Mobbs, "Differences in Behavior and Brain Activity during Hypothetical and Real Choices," *Trends Cogn Sci.*, 2017, *21* (1), 46–56.

Carlsson, Fredrik, Peter Frykblom, and Carl-Johan Lagerkvist, "Using cheap talk as a test of validity in choice experiments," *Economics Letters*, 2005, *89* (2), 147–152.

Champ, Patricia A. and Michael P. Welsh, *Survey Methodologies for Stated-Choice Studies*, Dordrecht: Springer Netherlands,

Chapman, Jonathan, Mark Dean, Pietro Ortoleva, Erik Snowberg, and Colin Camerer, "Willingness to Pay and Willingness to Accept are Probably Less Correlated Than You Think," Working Paper 23954, NBER October 2017.

Comerford, David A., "Response Bias in Survey Measures of Expectations: Evidence from the Survey of Consumer Expectationsâ Inflation Module," *Journal of Money, Credit and Banking*, 2023.

Corrigan, Jay R., Saleem Alhabash, Matthew Rousu, and Sean B. Cash, "How much is social media worth? Estimating the value of Facebook by paying users to stop using it," *PLOS ONE*, 12 2018, *13* (12), 1–11.

Cummings, Ronald G. and Laura O. Taylor, "Unbiased Value Estimates for Environmental Goods: A Cheap Talk Design for the Contingent Valuation Method," *American Economic Review*, June 1999, *89* (3), 649–665.

Davison, Erik, Yaqing Xiao, and Hongjun Yan, "Response Order Biases in Economic Surveys," December 2022. https://ssrn.com/abstract=3786894.

de Magistris, Tiziana, Azucena Gracia, and Rodolfo M. Nayga Jr., "On the Use of Honesty Priming Tasks to Mitigate Hypothetical Bias in Choice Experiments," *American Journal of Agricultural Economics*, 2013, *95* (5), 1136–1154.

De-Martino, B, D Kumaran, B Holt, and RJ. Dolan, "The neurobiology of reference-dependent value computation," *J Neurosci.*, 2009, *29* (12), 3833–42.

Ding, Min, Rajdeep Grewal, and John Liechty, "Incentive-Aligned Conjoint Analysis," *Journal of Marketing Research*, 2005, *42* (1), 67–82.

Flynn, Terry N., Jordan J. Louviere, Tim J. Peters, and Joanna Coast, "Best-worst scaling: What it can do for health care research and how to do it," *Journal of Health Economics*, 2007, *26* (1), 171–189.

_ , _ , _ , and _ , "Estimating preferences for a dermatology consultation using Best-Worst Scaling: Comparison of various methods of analysis," *BMC Medical Research Methodology*, 2008, *8* (1).

Frank, Morgan R., Manuel Cebrian, Galen Pickard, and Iyad Rahwan, "Validating Bayesian truth serum in large-scale online human experiments," *PLOS ONE*, 05 2017, *12* (5), 1–13.

Frederick, Shane, George Loewenstein, and Ted O'Donoghue, "Time discounting and time preference: A critical review," *Journal of Economic Literature*, 2002, *40* (2), 351 – 401. Cited by: 3281.

Haghani, Milad, Michiel C.J. Bliemer, John M. Rose, Harmen Oppewal, and Emily Lancsar, "Hypothetical bias in stated choice experiments: Part II. Conceptualisation of external validity, sources and explanations of bias and effectiveness of mitigation methods," *Journal of Choice Modelling*, 2021, *41*, 100322.

Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto, "Validating vignette and conjoint survey experiments against real-world behavior," *Proceedings of the National Academy of Sciences*, 2015, *112* (8), 2395–2400.

Hensher, D. A., J. M. Rose, and W. H. Greene, *Applied Choice Analysis*, Cambridge

University Press, 2015.

Jacquemet, Nicolas, Alexander G. James, Stephane Luchini, and Jason F. Shogren, "Social Psychology and Environmental Economics: A New Look at ex ante Corrections of Biased Preference Evaluation," *Environmental and Resource Economics*, 2011, *48*, 413–433.

_ , Alexander James, Stephane Luchini, and Jason F. Shogren, "Referenda Under Oath," *Environmental and Resource Economics*, 2017, *67*, 479 – 504.

Kasprzyk, Daniel, "Measurement error in household surveys: sources and measurement," December 2005. https://unstats.un.org/unsd/hhsurveys/pdf/chapter_9.pdf.

Konig, Pascal D, Stefan Wurster, and Markus B Siewert, "Consumers are willing to pay a price for explainable, but not for green AI. Evidence from a choice-based conjoint analysis," *Big Data & Society*, 2022, *9* (1), 20539517211069632.

Lee, Jinkwon and Uk Hwang, "Hypothetical Bias in Risk Preferences as a Driver of Hypothetical Bias in Willingness to Pay: Experimental Evidence," *Environmental Resource Economics*, 2016, *65* (4), 789–811.

Leggett, Christopher G., Naomi S. Kleckner, Kevin J. Boyle, John W. Dufield, and Robert Cameron Mitchell, "Social Desirability Bias in Contingent Valuation Surveys Administered Through In-Person Interviews," *Land Economics*, 2003, *79* (4), 561–575.

Lewis, Amy R., Richard P. Young, James M. Gibbons, and Julia P. G. Jones, "To what extent do potential conservation donors value community-aspects of conservation projects in low income countries?," *PLOS ONE*, 02 2018, *13* (2), 1–18.

List, John A., Paramita Sinha, and Michael H. Taylor, "Using Choice Experiments to Value Non-Market Goods and Services: Evidence from Field Experiments," *The B.E. Journal of Economic Analysis Policy*, 2006, *6* (2).

Loewenstein, George and David Schkade, "Wouldn't it be nice? Predicting future feelings," *Well-being: The foundations of hedonic psychology*, 1999, pp. 85–105.

_ , Ted O'Donoghue, and Matthew Rabin, "Projection bias in predicting future utility," *Quarterly Journal of Economics*, 2003, *118* (4), 1209 – 1248. Cited by: 549; All Open Access, Green Open Access.

Louviere, J. J., T. N. Flynn, and A. A. J. Marley, *Best-Worst Scaling: Theory, Methods and Applications*, Cambridge University Press, 2015.

Lusk, Jayson L., "Effects of Cheap Talk on Consumer Willingness-to-Pay for Golden Rice," *American Journal of Agricultural Economics*, 2003, *85* (4), 840–856.

_ , Deacue Fields, and Walt Prevatt, "An Incentive Compatible Conjoint Ranking Miechanism," *American Journal of Agricultural Economics*, 2008, *90* (2), 487–498.

Marley, A.A.J. and D. Pihlens, "Models of best-worst choice and ranking among multiattribute options (profiles)," *Journal of Mathematical Psychology*, 2012, *56* (1), 24–34.

Menapace, Luisa and Roberta Raffaelli, "Unraveling hypothetical bias in discrete choice experiments," *Journal of Economic Behavior Organization*, 2020, *176*, 416–430.

Moore, Jeffrey, Linda Stinson, and Edward Welniak, "Income Measurement Error in Surveys: A Review," June 1997. https://www.census.gov/content/dam/Census/library/working-papers/1997/adrm/sm97-05.pdf.

Morkbak, Morten Raun, Tove Christensen, and Dorte Gyrd-Hansen, "Choke Price Bias in Choice Experiments," *Environmental and Resource Economics*, 2014, *45.*

Mosquera, Roberto, Mofioluwasademi Odunowo, Trent McNamara, Xiongfei Guo, and Ragan Petrie, "The economic effects of Facebook," *Experimental Economics*, 2020, *23*, 575–602.

Murphy, J.J., P.G. Allen, and T.H. Stevens, "A Meta-analysis of Hypothetical Bias in Stated Preference Valuation," *Environ Resource Econ*, 2005, *30*, 313–325.

Nunes, Paulo A.L.D and Erik Schokkaert, "Identifying the warm glow effect in contingent valuation," *Journal of Environmental Economics and Management*, 2003, *45* (2), 231–245.

Olynk, Nicole J., Glynn T. Tonsor, and Christopher A. Wolf, "Consumer willingness to pay for livestock credence attribute claim verification," *Journal of Agricultural and Resource Economics*, 2010, *35* (2), 261 – 280. Cited by: 114.

Prelec, Drazen, "A Bayesian Truth Serum for Subjective Data," *Science*, 2004, *306* (5695), 462–466.

Sanjuan-Lopez, Ana I. and Helena Resano-Ezcaray, "Labels for a Local Food Speciality Product: The Case of Saffron," *Journal of Agricultural Economics*, 2020, *71* (3), 778–797.

Smith, Brett, Doina Olaru, Fakhra Jabeen, and Stephen Greaves, "Electric vehicles adoption: Environmental enthusiast bias in discrete choice models," *Transportation Research Part D: Transport and Environment*, 2017, *51*, 290–303.

Svenningsen, Lea S. and Jette Bredahl Jacobsen, "Testing the effect of changes in elicitation format, payment vehicle and bid range on the hypothetical bias for moral goods," *Journal of Choice Modelling*, 2018, *29*, 17–32.

Weaver, Ray and Drazen Prelec, "Creating Truth-Telling Incentives with the Bayesian Truth Serum," *Journal of Marketing Research*, 2013, *50* (3), 289–302.

Wegener, D. T. and R. E. Petty, "Flexible correction processes in social judgment: The role of naive theories in corrections for perceived bias," *Journal of Personality and Social Psychology*, 1995, *68* (1), 36–51.

Whittington, Dale, V.Kerry Smith, Apia Okorafor, Augustine Okore, Jin Long Liu, and Alexander McPhail, "Giving respondents time to think in contingent valuation studies: A developing country application," *Journal of Environmental Economics and Management*, 1992, *22* (3), 205–225.

Wittenberg, E, M Bharel, JF Bridges, Z Ward, and L Weinreb, "Using Best-Worst Scaling to Understand Patient Priorities: A Case Example of Papanicolaou Tests for Homeless Women," *Ann Fam Med.*, July 2016, *14* (4), 359–64.

Zhang, Hao, Xiaofei Bai, and Zengguang Ma, "Consumer reactions to AI design: Exploring consumer willingness to pay for AI-designed products," *Psychology & Marketing*, 2022, *39* (11), 2171–2183.

Online Appendix

Cheap Talk with the Bayesian Truth Serum

A Additional Results

Figure Appendix 1: The Results on Donation Decision-making in Barrage and Lee (2010)

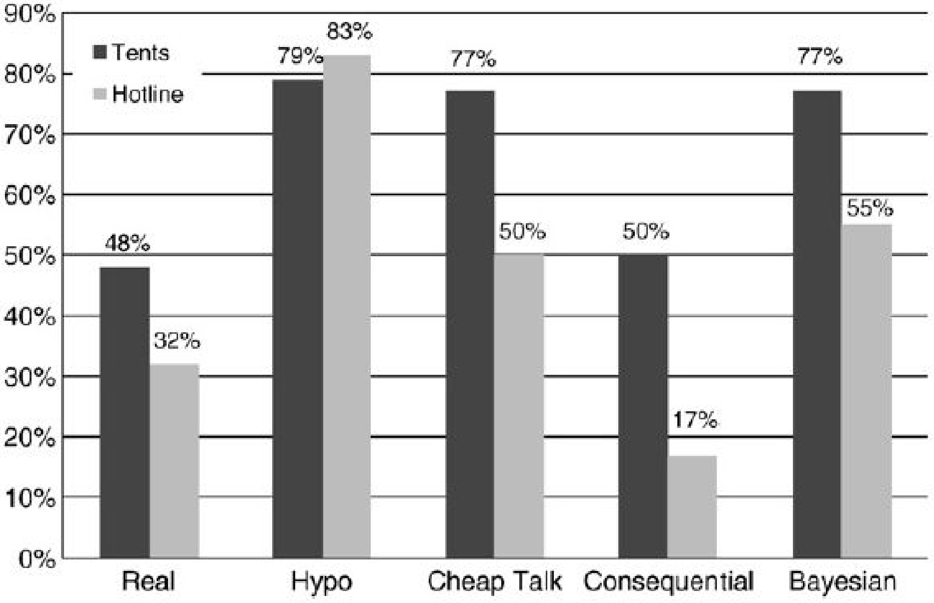


Figure Appendix 2: The Full Sample Results on Donation Deicision-Making without Attention Check

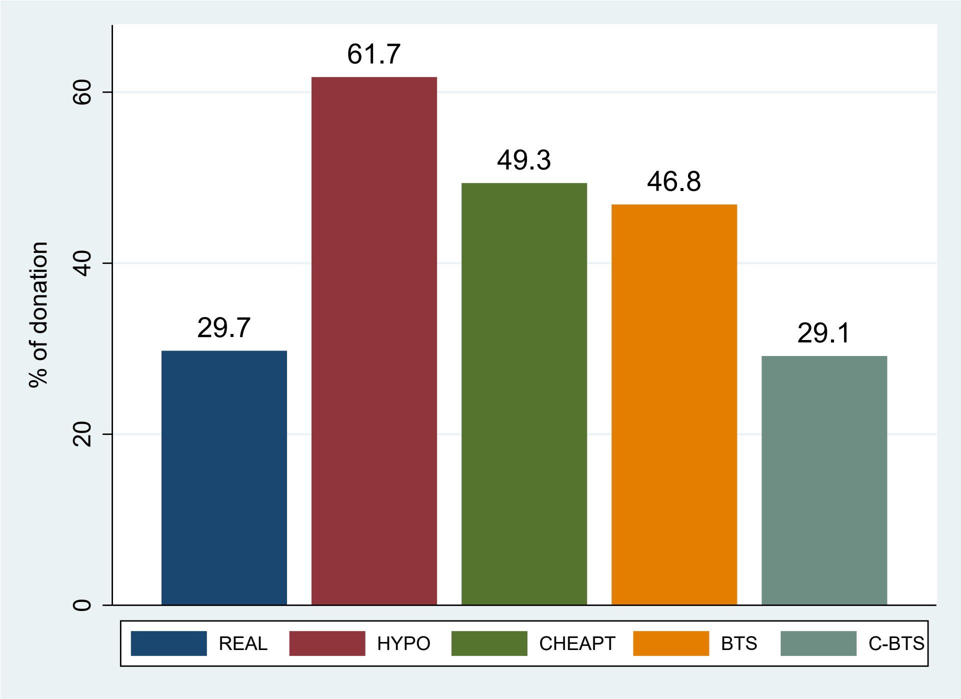
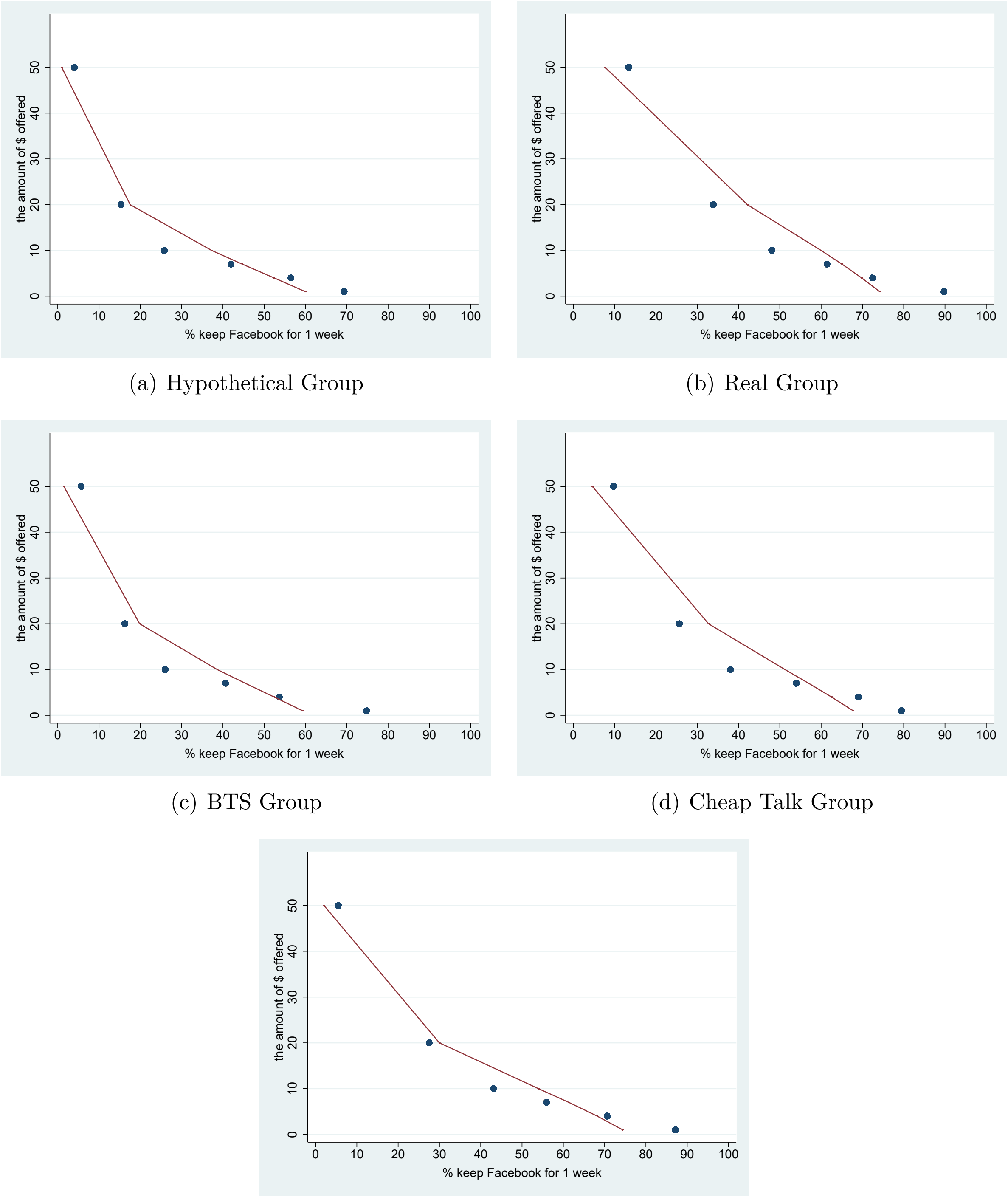


Figure Appendix 3: The Demand Curve for Facebook from Binary Discrete Choices



(a) Hypothetical Group

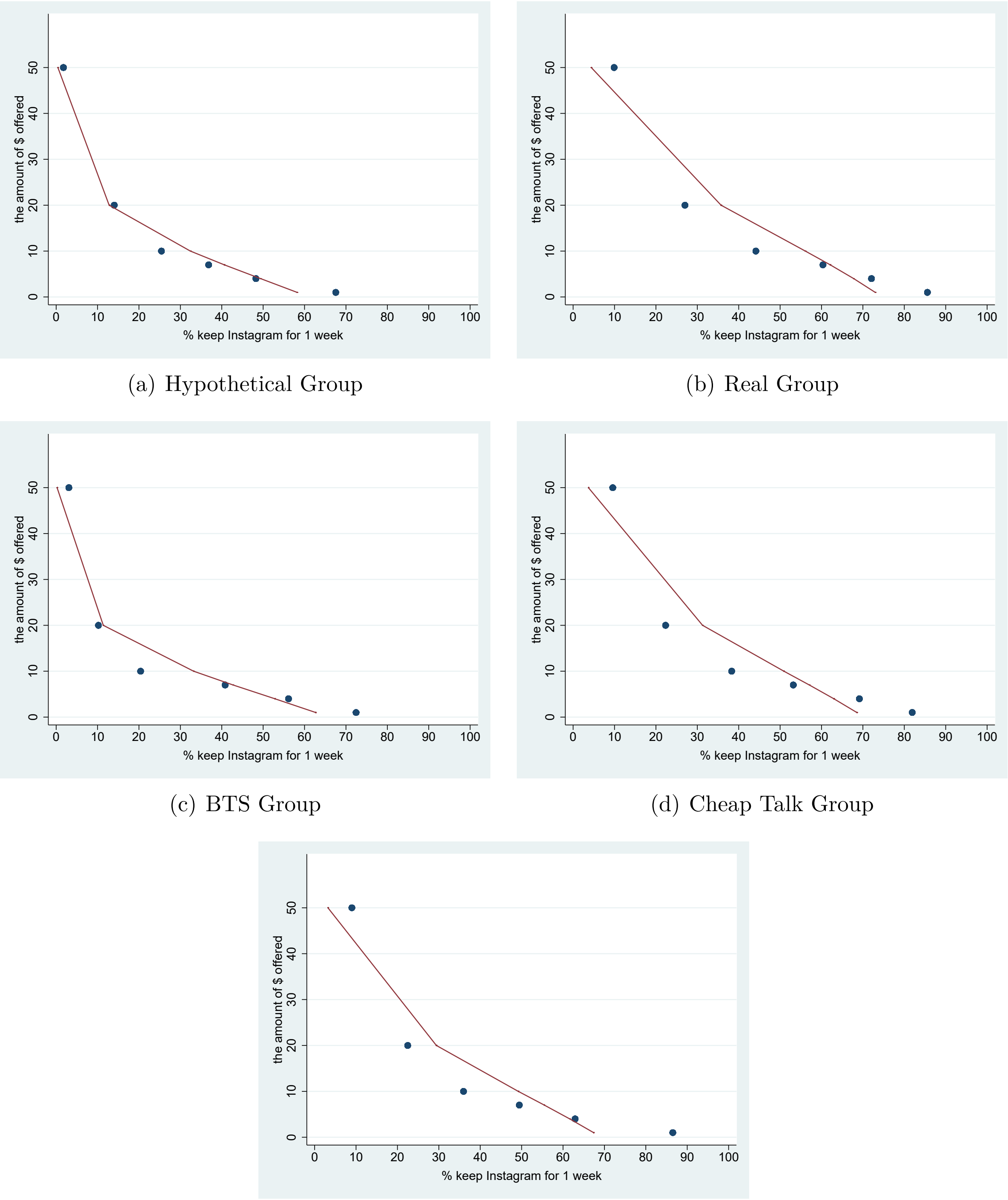
(b) Real Group

(c) BTS Group

(d) Cheap Talk Group

(e) C-BTS Group

Figure Appendix 4: The Demand Curve for Instagram from Binary Discrete Choices



(a) Hypothetical Group

(b) Real Group

(c) BTS Group

(d) Cheap Talk Group

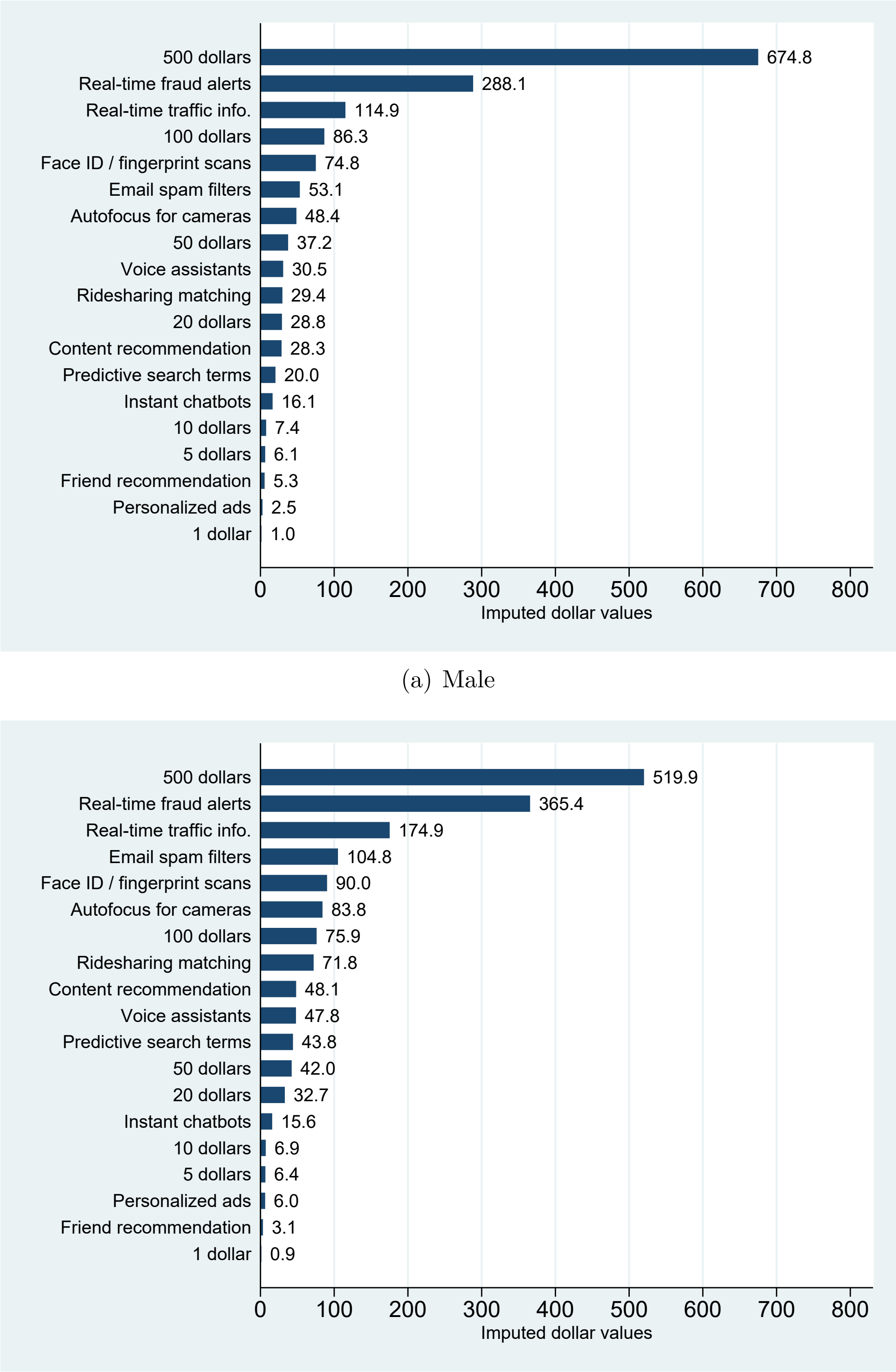
(e) C-BTS Group

Table Appendix 1: The Conditional Logit Estimates for the Value of AI-powered Services

AI-powered services	coefficients	standard errors
Fraud alerts	-3.760	(0.129)
Real-time traffic	-3.242	(0.121)
Email spam filters	-2.834	(0.117)
Search terms	-2.243	(0.114)
Friend recomm.	-0.913	(0.125)
Personal ads.	-0.952	(0.120)
Voice assistants	-2.400	(0.117)
Face ID	-2.878	(0.124)
Autofocus	-2.728	(0.116)
Content recomm.	-2.375	(0.115)
Chatbot	-1.808	(0.117)
Ridesharing	-2.502	(0.132)
Earning \$5 less	-1.213	(0.113)
Earning \$10 less	-1.292	(0.100)
Earning \$20 less	-2.231	(0.106)
Earning \$50 less	-2.399	(0.110)
Earning \$100 less	-2.843	(0.112)
Earning \$500 less	-4.109	(0.125)
Earning \$1,000 less	-4.486	(0.132)
Earning \$5,000 less	-6.402	(0.191)
Earning \$1 less	-	-
Observations	25,710	

* The standard errors of the estimated coefficients are given in parentheses.

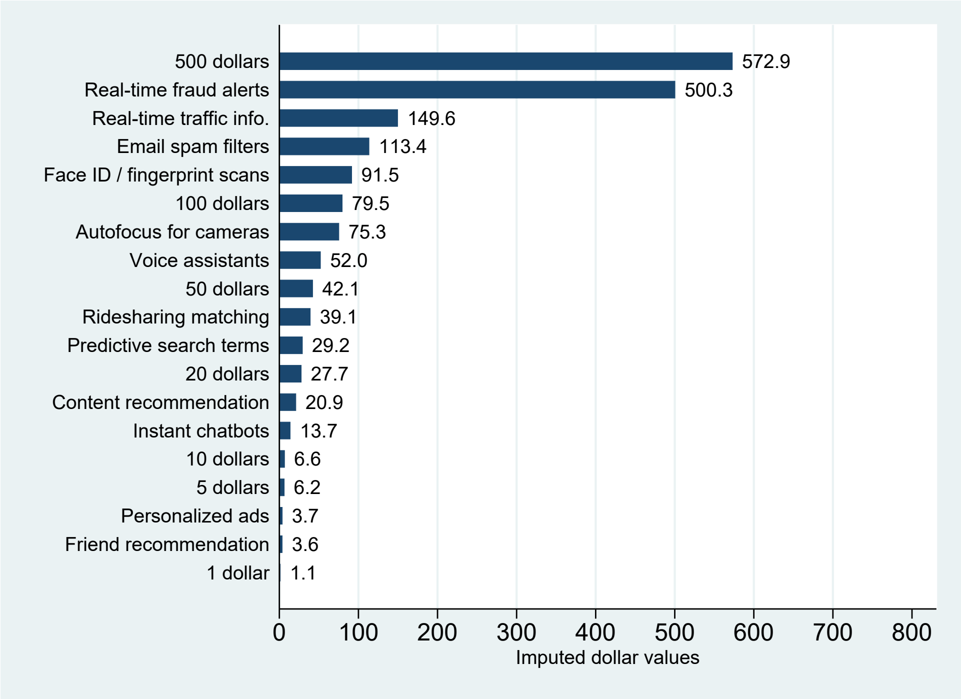
Figure Appendix 5: The Consumer Value of the AI-powered Services by Gender



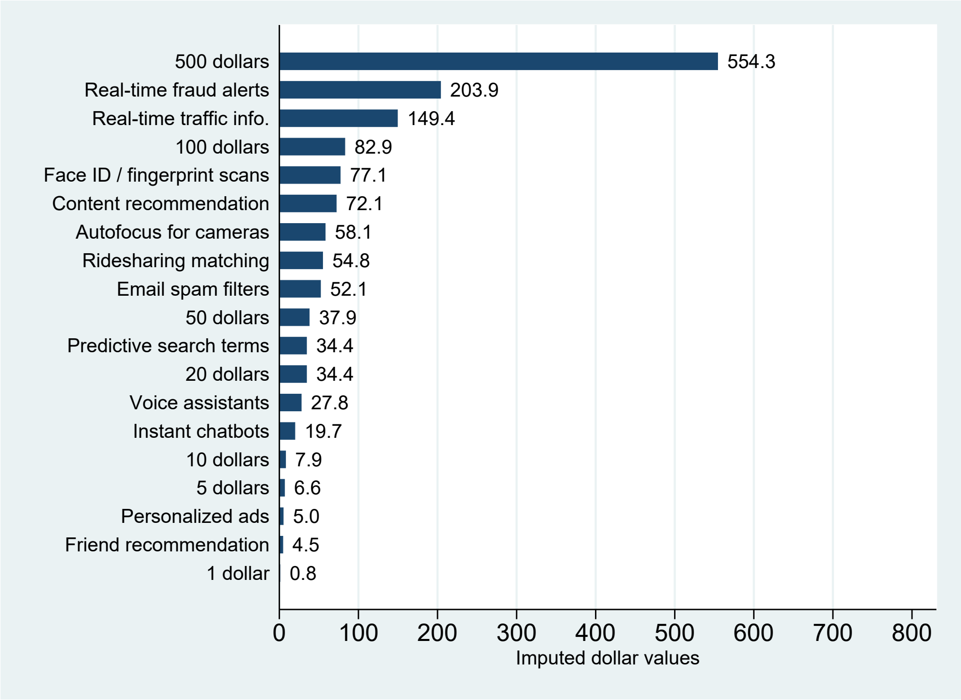
(a) Male

(b) Female

Figure Appendix 6: The Consumer Value of the AI-powered Services by Age

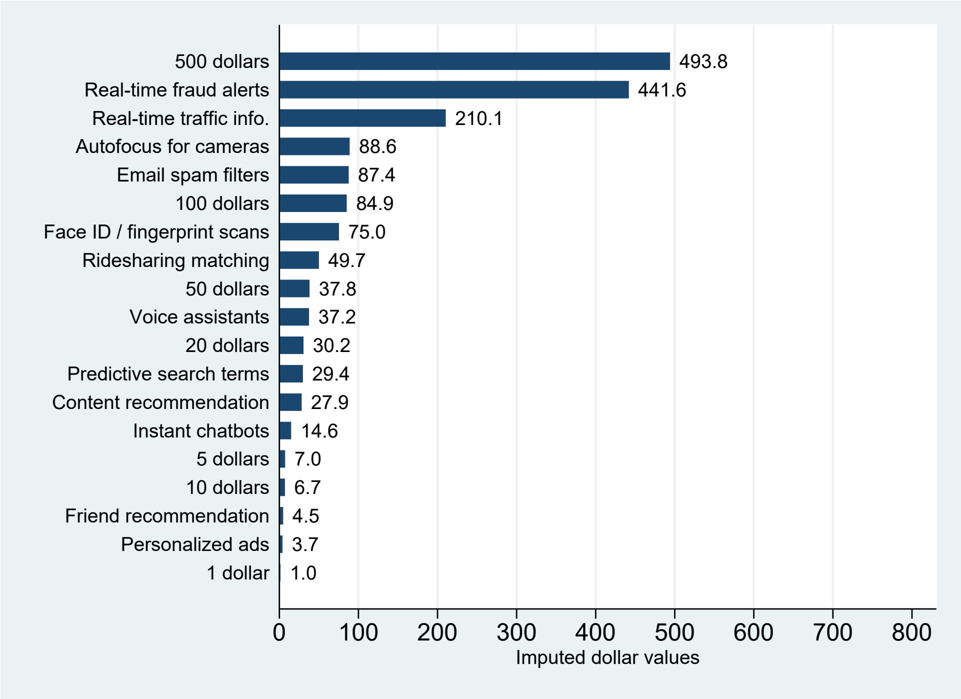


(a) Old

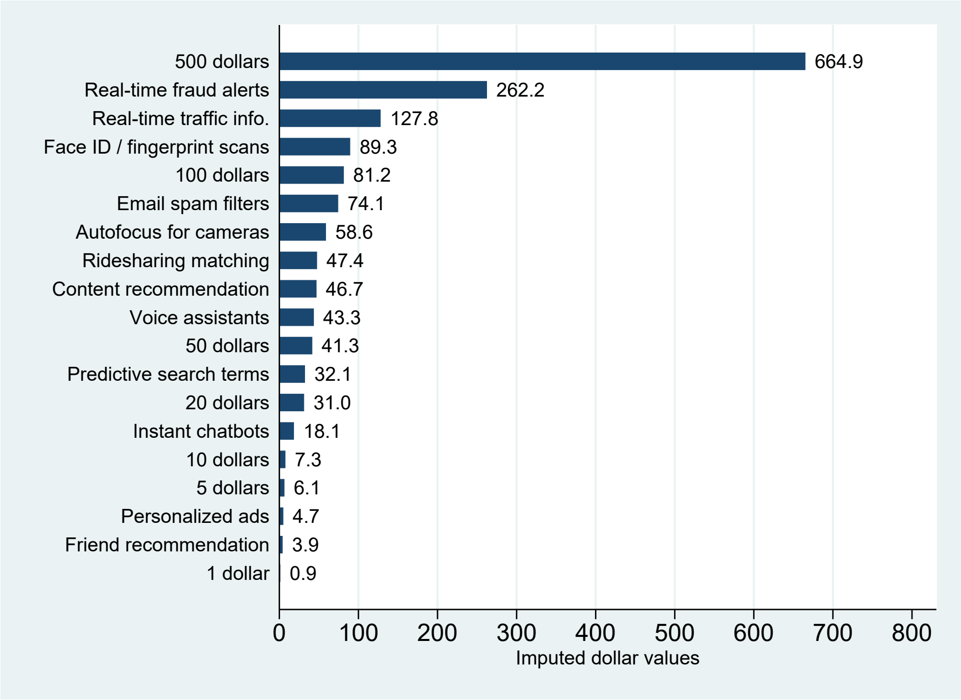


(b) Young

Figure Appendix 7: The Consumer Value of the AI-powered Services by Income

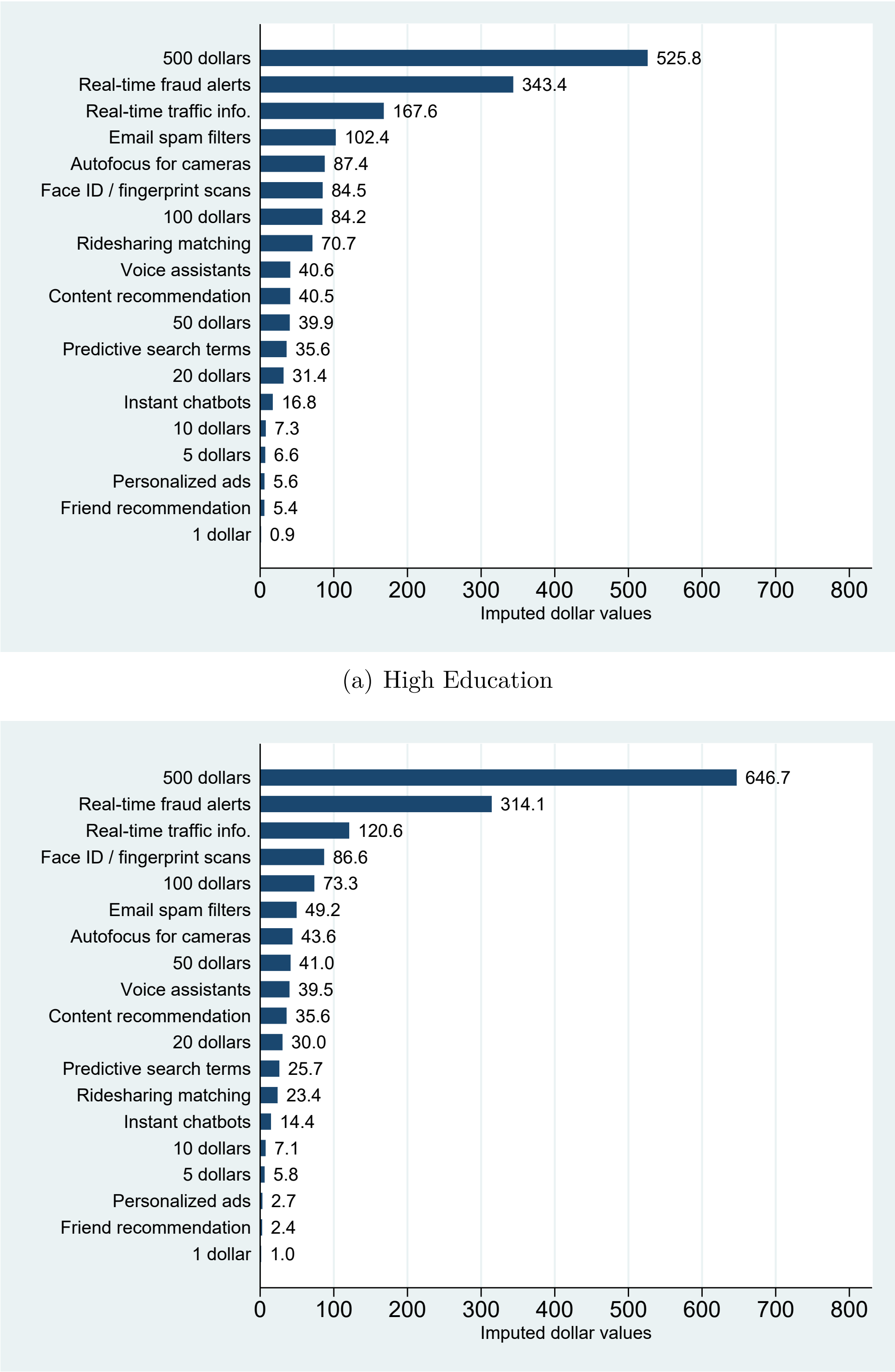


(a) High Income



(b) Low Income

Figure Appendix 8: The Consumer Value of the AI-powered Services by Education

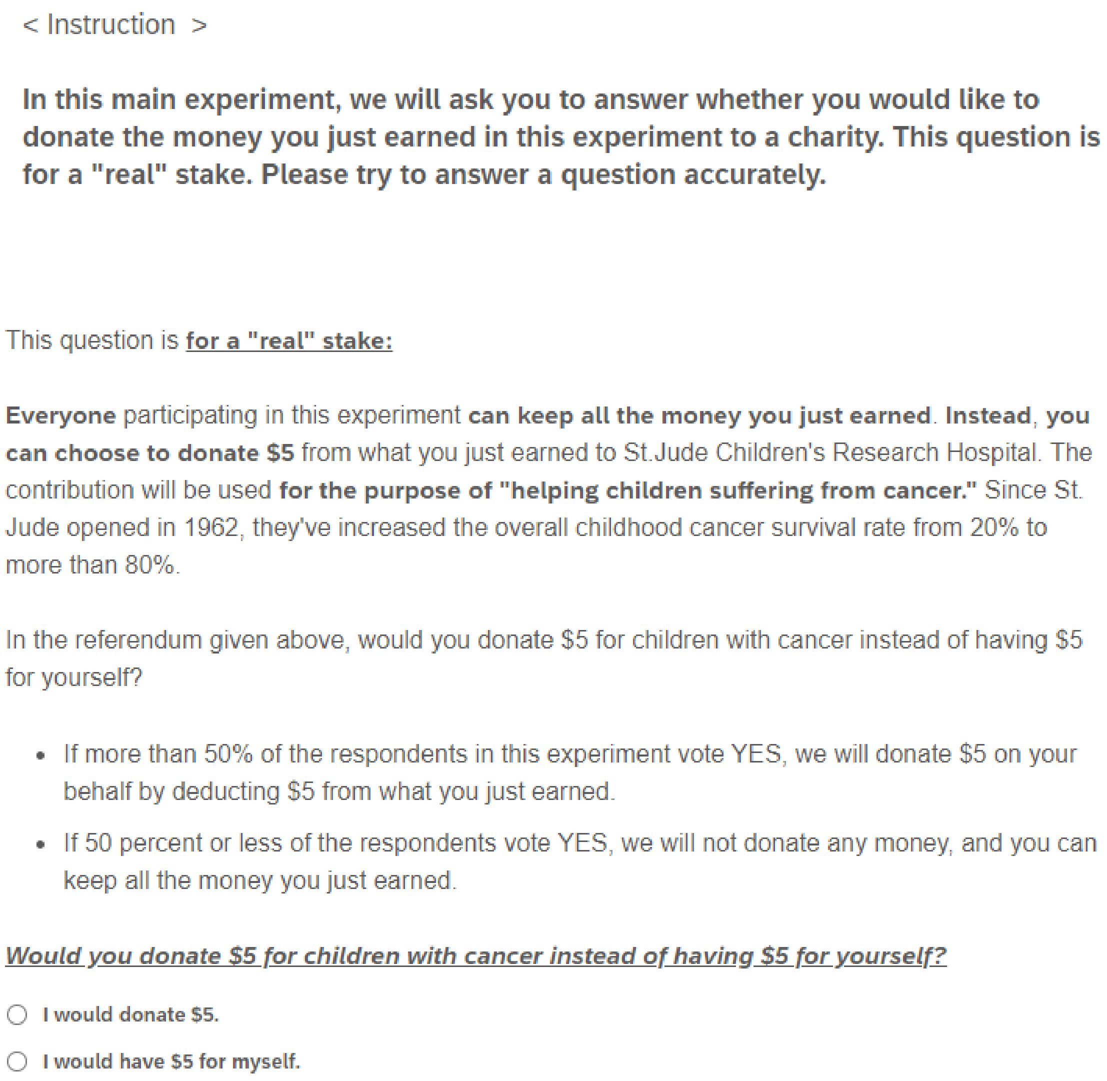


(a) High Education

(b) Low Education

B Instructions and Sample Questions

Figure Appendix 9: Real Group in Donation Decision-Making



< Instruction >

In this main experiment, we will ask you to answer whether you would like to donate the money you just earned in this experiment to a charity. This question is for a "real" stake. Please try to answer a question accurately.

This question is **for a "real" stake:**

Everyone participating in this experiment **can keep all the money you just earned**. **Instead**, **you can choose to donate $5** from what you just earned to St.Jude Children's Research Hospital. The contribution will be used **for the purpose of "helping children suffering from cancer."** Since St. Jude opened in 1962, they've increased the overall childhood cancer survival rate from 20% to more than 80%.

In the referendum given above, would you donate $5 for children with cancer instead of having $5 for yourself?

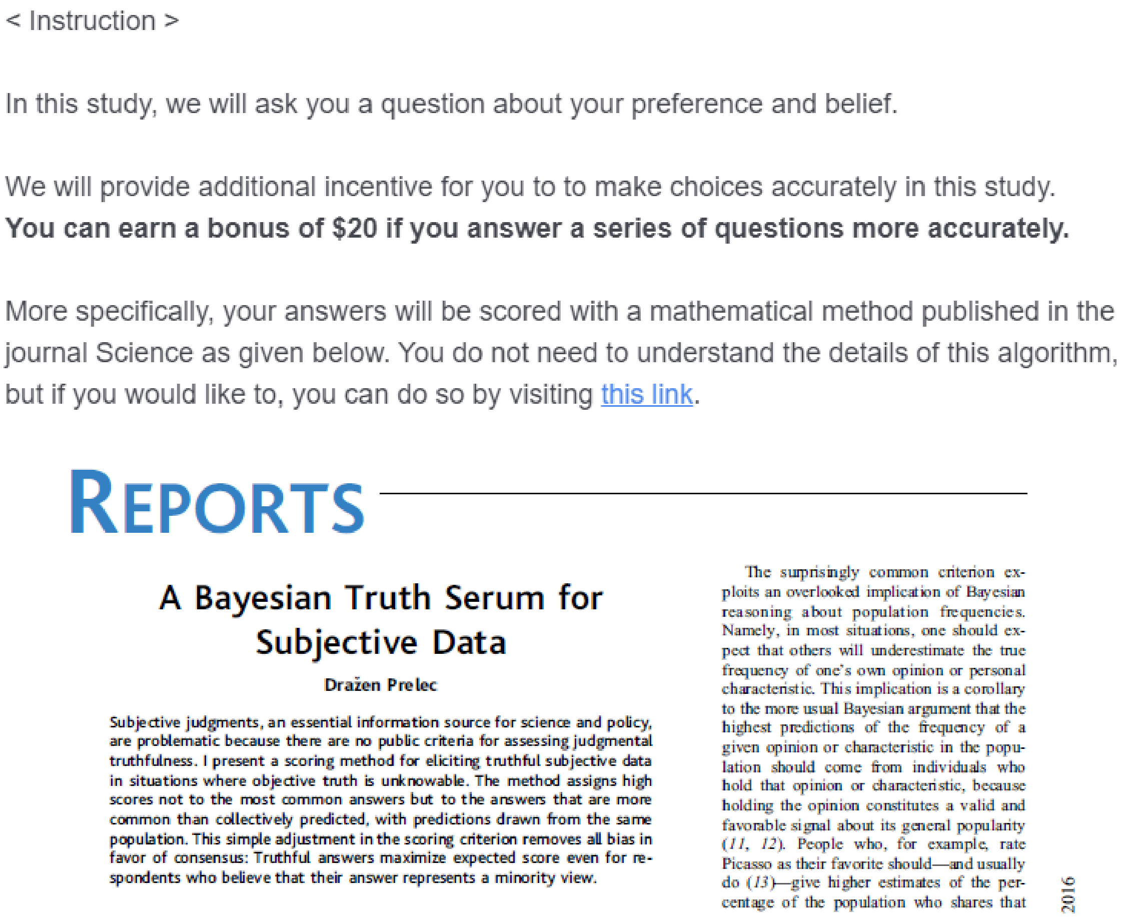
- If more than 50% of the respondents in this experiment vote YES, we will donate $5 on your behalf by deducting $5 from what you just earned.
- If 50 percent or less of the respondents vote YES, we will not donate any money, and you can keep all the money you just earned.

Would you donate $5 for children with cancer instead of having $5 for yourself?

○ **I would donate $5.**

○ **I would have $5 for myself.**

Figure Appendix 10: C-BTS Group in Donation Decision-Making



< Instruction >

In this study, we will ask you a question about your preference and belief.

We will provide additional incentive for you to to make choices accurately in this study.
You can earn a bonus of $20 if you answer a series of questions more accurately.

More specifically, your answers will be scored with a mathematical method published in the journal Science as given below. You do not need to understand the details of this algorithm, but if you would like to, you can do so by visiting this link.

REPORTS

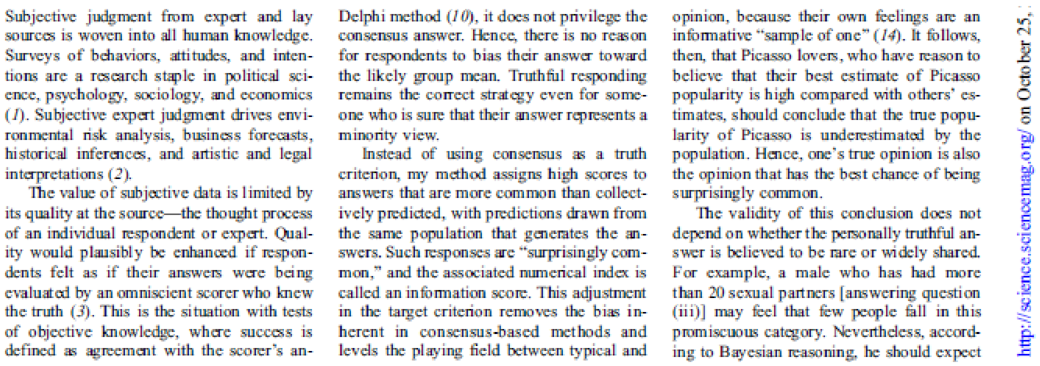
A Bayesian Truth Serum for Subjective Data

Dražen Prelec

Subjective judgments, an essential information source for science and policy, are problematic because there are no public criteria for assessing judgmental truthfulness. I present a scoring method for eliciting truthful subjective data in situations where objective truth is unknowable. The method assigns high scores not to the most common answers but to the answers that are more common than collectively predicted, with predictions drawn from the same population. This simple adjustment in the scoring criterion removes all bias in favor of consensus: Truthful answers maximize expected score even for respondents who believe that their answer represents a minority view.

The surprisingly common criterion exploits an overlooked implication of Bayesian reasoning about population frequencies. Namely, in most situations, one should expect that others will underestimate the true frequency of one's own opinion or personal characteristic. This implication is a corollary to the more usual Bayesian argument that the highest predictions of the frequency of a given opinion or characteristic in the population should come from individuals who hold that opinion or characteristic, because holding the opinion constitutes a valid and favorable signal about its general popularity (*11*, *12*). People who, for example, rate Picasso as their favorite should—and usually do (*13*)—give higher estimates of the percentage of the population who shares that

2016



Subjective judgment from expert and lay sources is woven into all human knowledge. Surveys of behaviors, attitudes, and intentions are a research staple in political science, psychology, sociology, and economics (*1*). Subjective expert judgment drives environmental risk analysis, business forecasts, historical inferences, and artistic and legal interpretations (*2*).

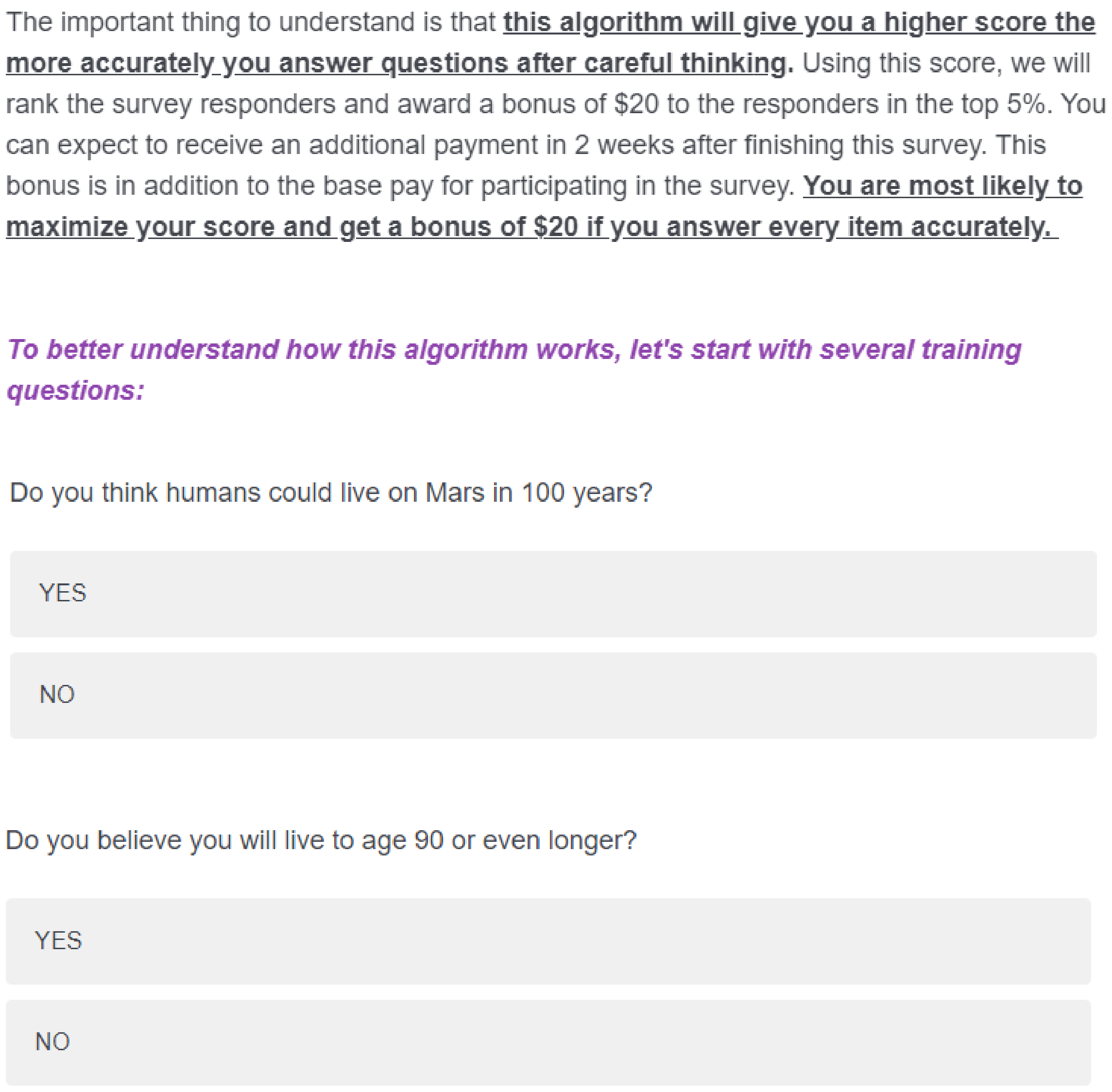
The value of subjective data is limited by its quality at the source—the thought process of an individual respondent or expert. Quality would plausibly be enhanced if respondents felt as if their answers were being evaluated by an omniscient scorer who knew the truth (*3*). This is the situation with tests of objective knowledge, where success is defined as agreement with the scorer's an-

Delphi method (*10*), it does not privilege the consensus answer. Hence, there is no reason for respondents to bias their answer toward the likely group mean. Truthful responding remains the correct strategy even for someone who is sure that their answer represents a minority view.

Instead of using consensus as a truth criterion, my method assigns high scores to answers that are more common than collectively predicted, with predictions drawn from the same population that generates the answers. Such responses are "surprisingly common," and the associated numerical index is called an information score. This adjustment in the target criterion removes the bias inherent in consensus-based methods and levels the playing field between typical and

opinion, because their own feelings are an informative "sample of one" (*14*). It follows, then, that Picasso lovers, who have reason to believe that their best estimate of Picasso popularity is high compared with others' estimates, should conclude that the true popularity of Picasso is underestimated by the population. Hence, one's true opinion is also the opinion that has the best chance of being surprisingly common.

The validity of this conclusion does not depend on whether the personally truthful answer is believed to be rare or widely shared. For example, a male who has had more than 20 sexual partners [answering question (iii)] may feel that few people fall in this promiscuous category. Nevertheless, according to Bayesian reasoning, he should expect



The important thing to understand is that **this algorithm will give you a higher score the more accurately you answer questions after careful thinking.** Using this score, we will rank the survey responders and award a bonus of $20 to the responders in the top 5%. You can expect to receive an additional payment in 2 weeks after finishing this survey. This bonus is in addition to the base pay for participating in the survey. **You are most likely to maximize your score and get a bonus of $20 if you answer every item accurately.**

To better understand how this algorithm works, let's start with several training questions:

Do you think humans could live on Mars in 100 years?

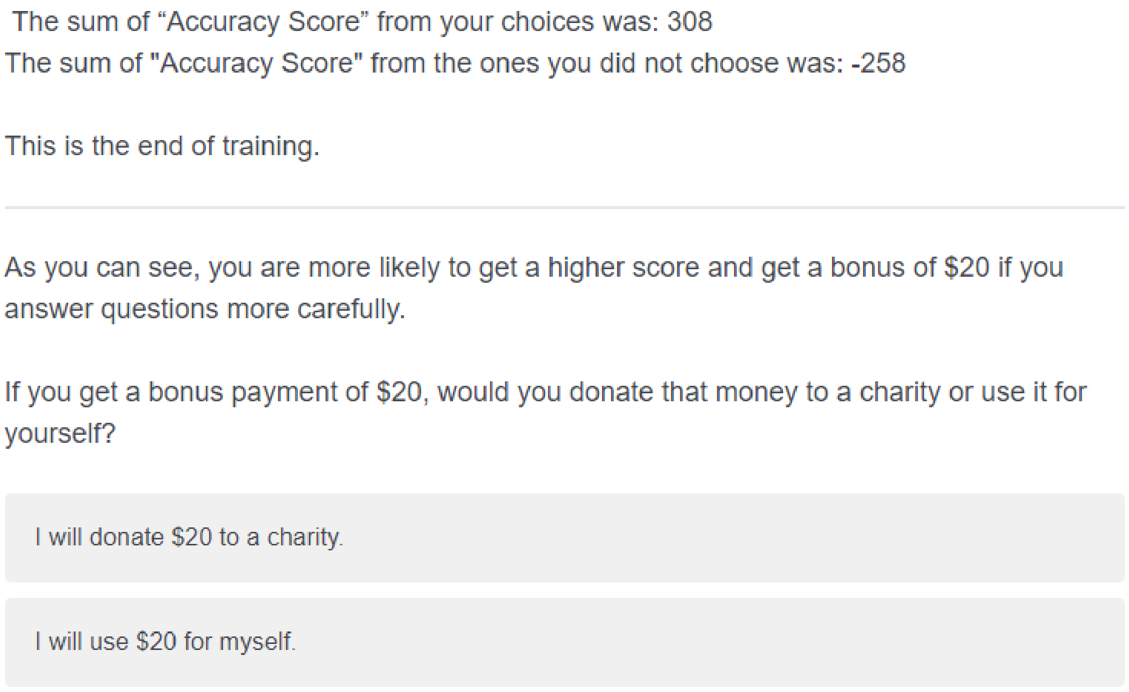
YES

NO

Do you believe you will live to age 90 or even longer?

YES

NO



The sum of “Accuracy Score” from your choices was: 308
The sum of "Accuracy Score" from the ones you did not choose was: -258

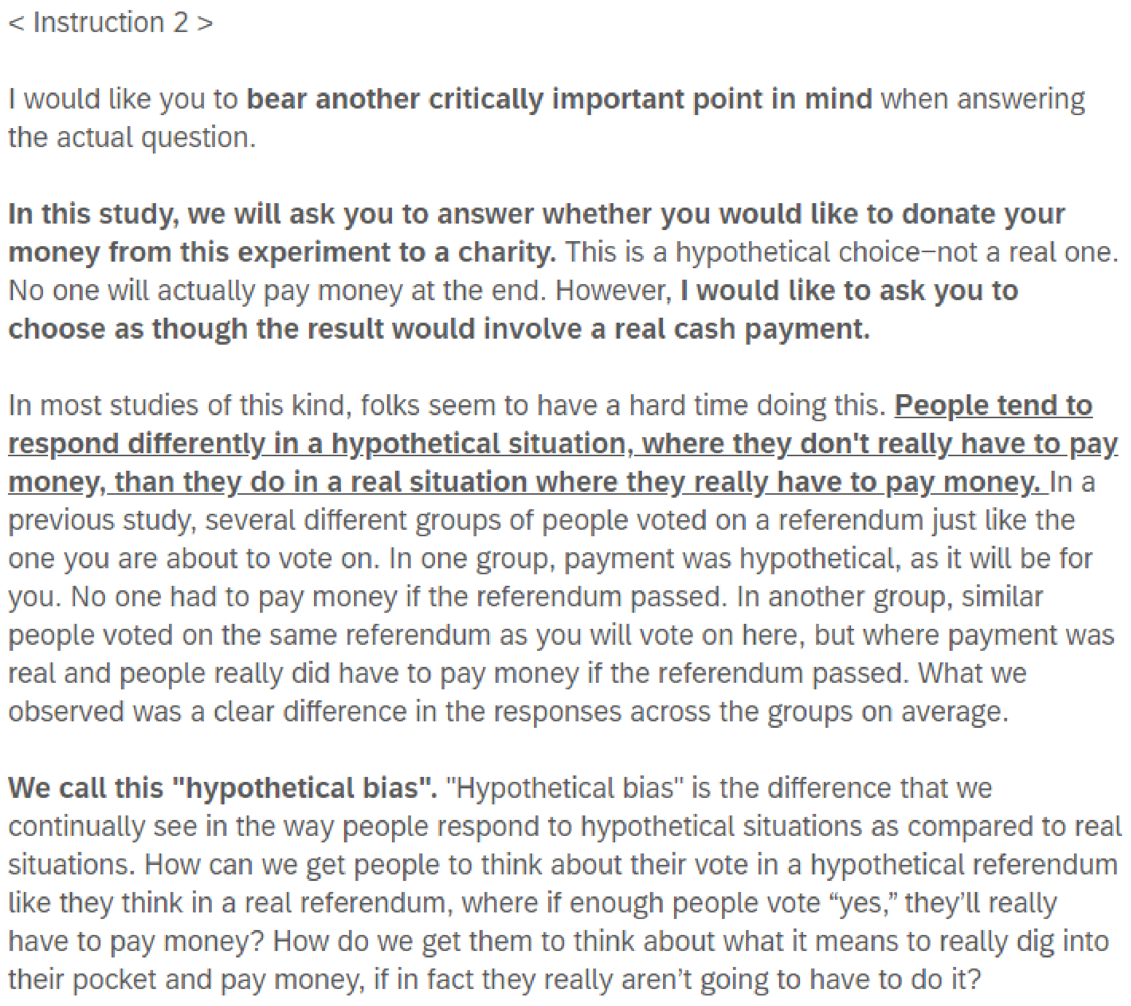
This is the end of training.

As you can see, you are more likely to get a higher score and get a bonus of $20 if you answer questions more carefully.

If you get a bonus payment of $20, would you donate that money to a charity or use it for yourself?

I will donate $20 to a charity.

I will use $20 for myself.



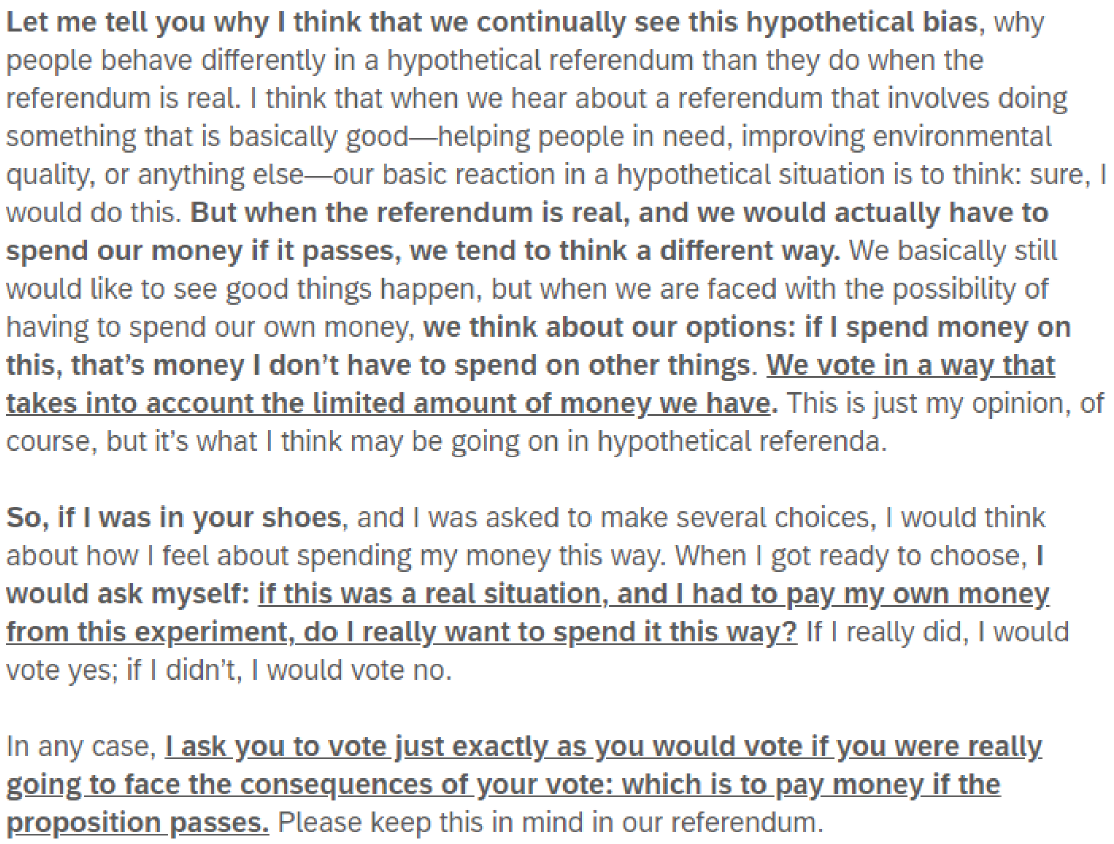
< Instruction 2 >

I would like you to **bear another critically important point in mind** when answering the actual question.

In this study, we will ask you to answer whether you would like to donate your money from this experiment to a charity. This is a hypothetical choice–not a real one. No one will actually pay money at the end. However, **I would like to ask you to choose as though the result would involve a real cash payment.**

In most studies of this kind, folks seem to have a hard time doing this. **People tend to respond differently in a hypothetical situation, where they don't really have to pay money, than they do in a real situation where they really have to pay money.** In a previous study, several different groups of people voted on a referendum just like the one you are about to vote on. In one group, payment was hypothetical, as it will be for you. No one had to pay money if the referendum passed. In another group, similar people voted on the same referendum as you will vote on here, but where payment was real and people really did have to pay money if the referendum passed. What we observed was a clear difference in the responses across the groups on average.

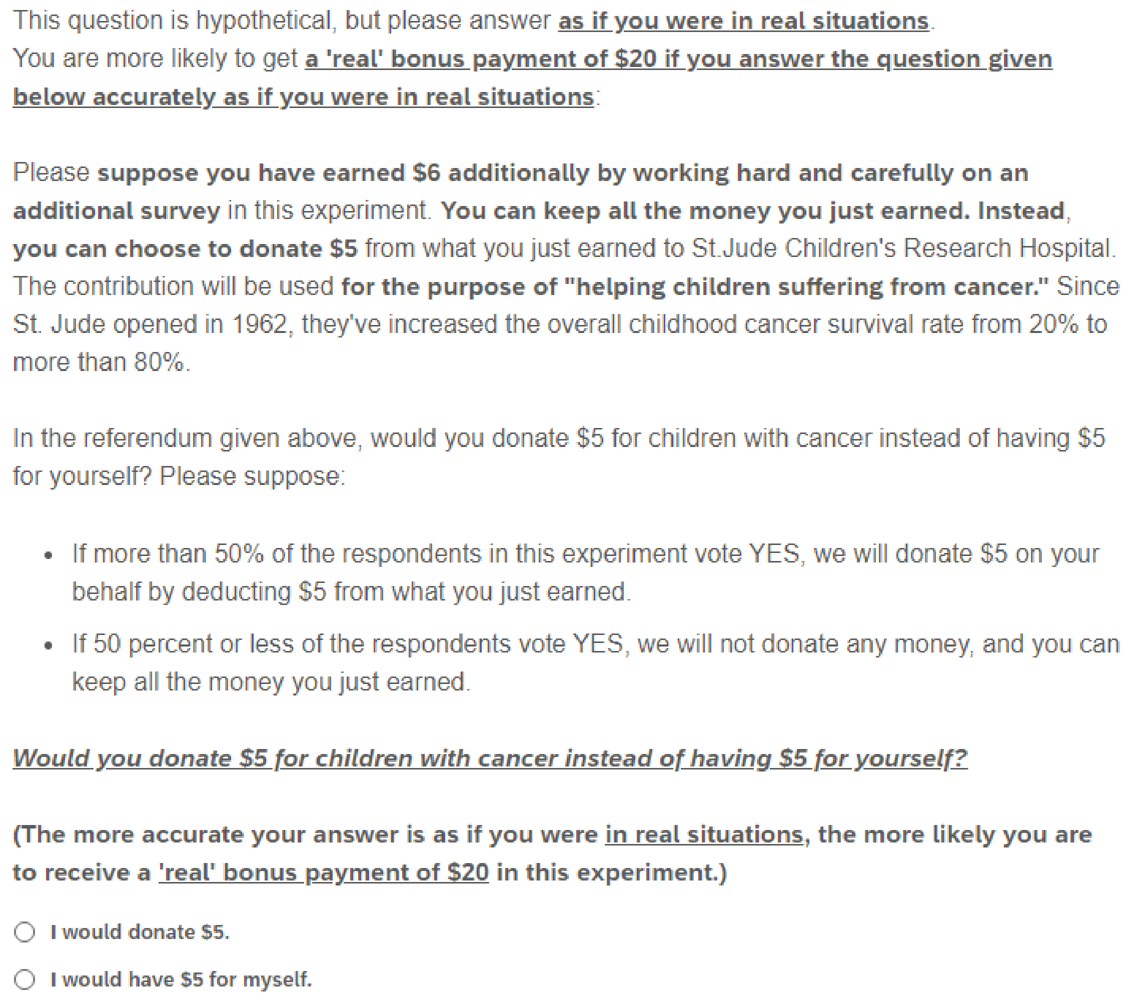
We call this "hypothetical bias". "Hypothetical bias" is the difference that we continually see in the way people respond to hypothetical situations as compared to real situations. How can we get people to think about their vote in a hypothetical referendum like they think in a real referendum, where if enough people vote “yes,” they’ll really have to pay money? How do we get them to think about what it means to really dig into their pocket and pay money, if in fact they really aren’t going to have to do it?



Let me tell you why I think that we continually see this hypothetical bias, why people behave differently in a hypothetical referendum than they do when the referendum is real. I think that when we hear about a referendum that involves doing something that is basically good—helping people in need, improving environmental quality, or anything else—our basic reaction in a hypothetical situation is to think: sure, I would do this. **But when the referendum is real, and we would actually have to spend our money if it passes, we tend to think a different way.** We basically still would like to see good things happen, but when we are faced with the possibility of having to spend our own money, **we think about our options: if I spend money on this, that's money I don't have to spend on other things**. **We vote in a way that takes into account the limited amount of money we have.** This is just my opinion, of course, but it's what I think may be going on in hypothetical referenda.

So, if I was in your shoes, and I was asked to make several choices, I would think about how I feel about spending my money this way. When I got ready to choose, **I would ask myself: if this was a real situation, and I had to pay my own money from this experiment, do I really want to spend it this way?** If I really did, I would vote yes; if I didn't, I would vote no.

In any case, **I ask you to vote just exactly as you would vote if you were really going to face the consequences of your vote: which is to pay money if the proposition passes.** Please keep this in mind in our referendum.



This question is hypothetical, but please answer **as if you were in real situations**.
You are more likely to get **a 'real' bonus payment of $20 if you answer the question given below accurately as if you were in real situations**:

Please **suppose you have earned $6 additionally by working hard and carefully on an additional survey** in this experiment. **You can keep all the money you just earned. Instead, you can choose to donate $5** from what you just earned to St.Jude Children's Research Hospital. The contribution will be used **for the purpose of "helping children suffering from cancer."** Since St. Jude opened in 1962, they've increased the overall childhood cancer survival rate from 20% to more than 80%.

In the referendum given above, would you donate $5 for children with cancer instead of having $5 for yourself? Please suppose:

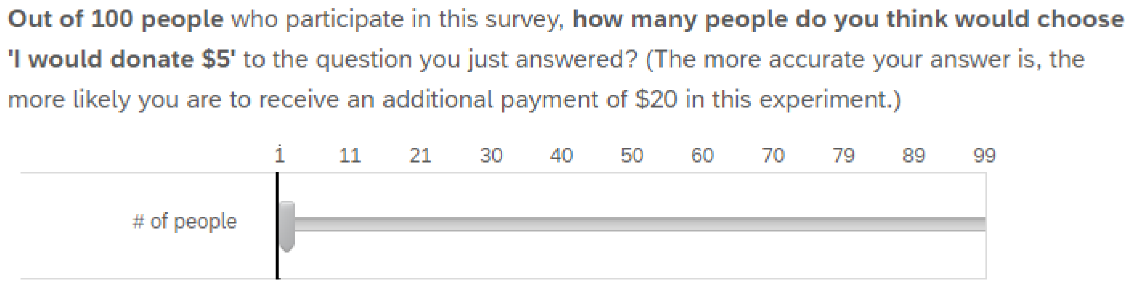
- If more than 50% of the respondents in this experiment vote YES, we will donate $5 on your behalf by deducting $5 from what you just earned.
- If 50 percent or less of the respondents vote YES, we will not donate any money, and you can keep all the money you just earned.

Would you donate $5 for children with cancer instead of having $5 for yourself?

(The more accurate your answer is as if you were in real situations, the more likely you are to receive a 'real' bonus payment of $20 in this experiment.)

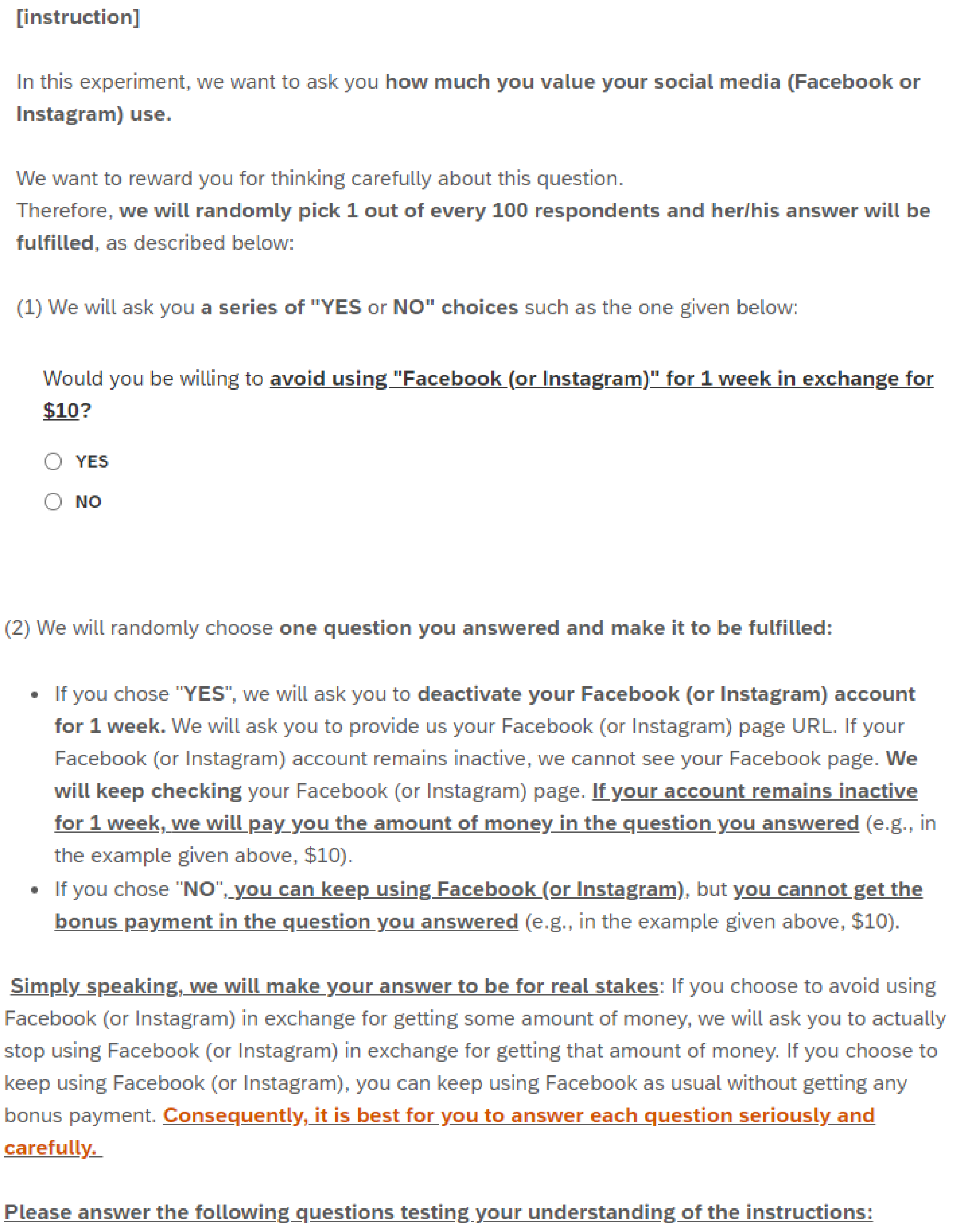
○ **I would donate $5.**

○ **I would have $5 for myself.**



Out of 100 people who participate in this survey, **how many people do you think would choose 'I would donate $5'** to the question you just answered? (The more accurate your answer is, the more likely you are to receive an additional payment of $20 in this experiment.)

Figure Appendix 11: Real Group in Measuring the WTA Using Binary Discrete Choices



[instruction]

In this experiment, we want to ask you **how much you value your social media (Facebook or Instagram) use.**

We want to reward you for thinking carefully about this question.
Therefore, **we will randomly pick 1 out of every 100 respondents and her/his answer will be fulfilled**, as described below:

(1) We will ask you **a series of "YES** or **NO" choices** such as the one given below:

Would you be willing to **avoid using "Facebook (or Instagram)" for 1 week in exchange for $10?**

○ **YES**

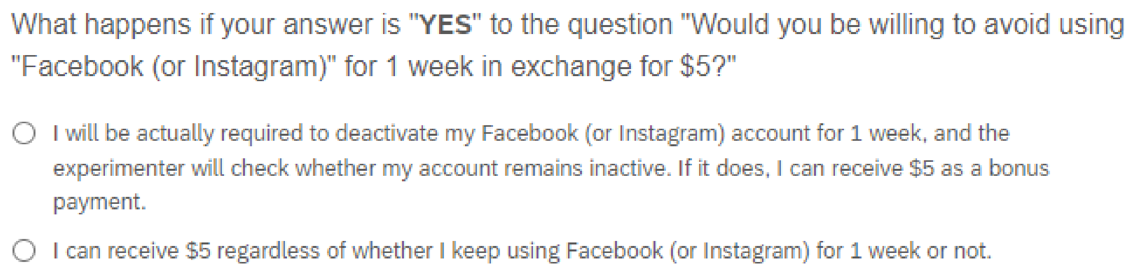
○ **NO**

(2) We will randomly choose **one question you answered and make it to be fulfilled:**

- If you chose "**YES**", we will ask you to **deactivate your Facebook (or Instagram) account for 1 week.** We will ask you to provide us your Facebook (or Instagram) page URL. If your Facebook (or Instagram) account remains inactive, we cannot see your Facebook page. **We will keep checking** your Facebook (or Instagram) page. **If your account remains inactive for 1 week, we will pay you the amount of money in the question you answered** (e.g., in the example given above, $10).
- If you chose "**NO**", **you can keep using Facebook (or Instagram)**, but **you cannot get the bonus payment in the question you answered** (e.g., in the example given above, $10).

Simply speaking, we will make your answer to be for real stakes: If you choose to avoid using Facebook (or Instagram) in exchange for getting some amount of money, we will ask you to actually stop using Facebook (or Instagram) in exchange for getting that amount of money. If you choose to keep using Facebook (or Instagram), you can keep using Facebook as usual without getting any bonus payment. **Consequently, it is best for you to answer each question seriously and carefully.**

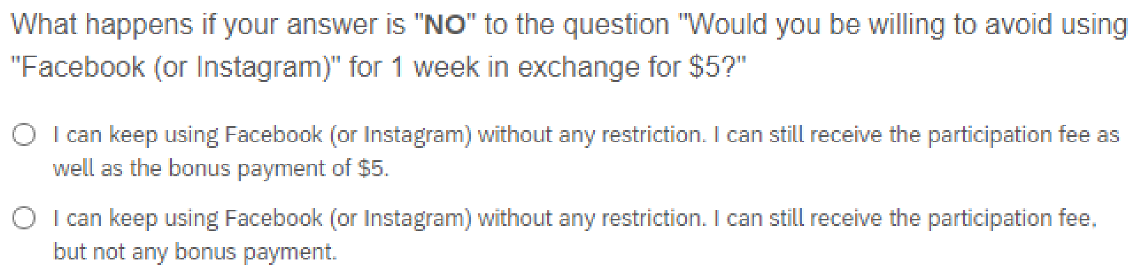
Please answer the following questions testing your understanding of the instructions:



What happens if your answer is "**YES**" to the question "Would you be willing to avoid using "Facebook (or Instagram)" for 1 week in exchange for $5?"

○ I will be actually required to deactivate my Facebook (or Instagram) account for 1 week, and the experimenter will check whether my account remains inactive. If it does, I can receive $5 as a bonus payment.

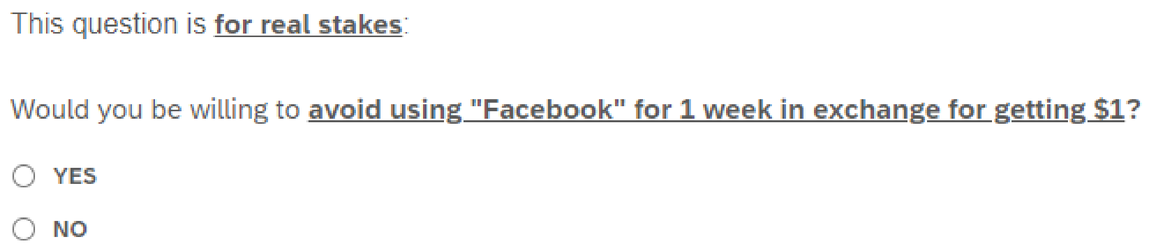
○ I can receive $5 regardless of whether I keep using Facebook (or Instagram) for 1 week or not.



What happens if your answer is "**NO**" to the question "Would you be willing to avoid using "Facebook (or Instagram)" for 1 week in exchange for $5?"

○ I can keep using Facebook (or Instagram) without any restriction. I can still receive the participation fee as well as the bonus payment of $5.

○ I can keep using Facebook (or Instagram) without any restriction. I can still receive the participation fee, but not any bonus payment.



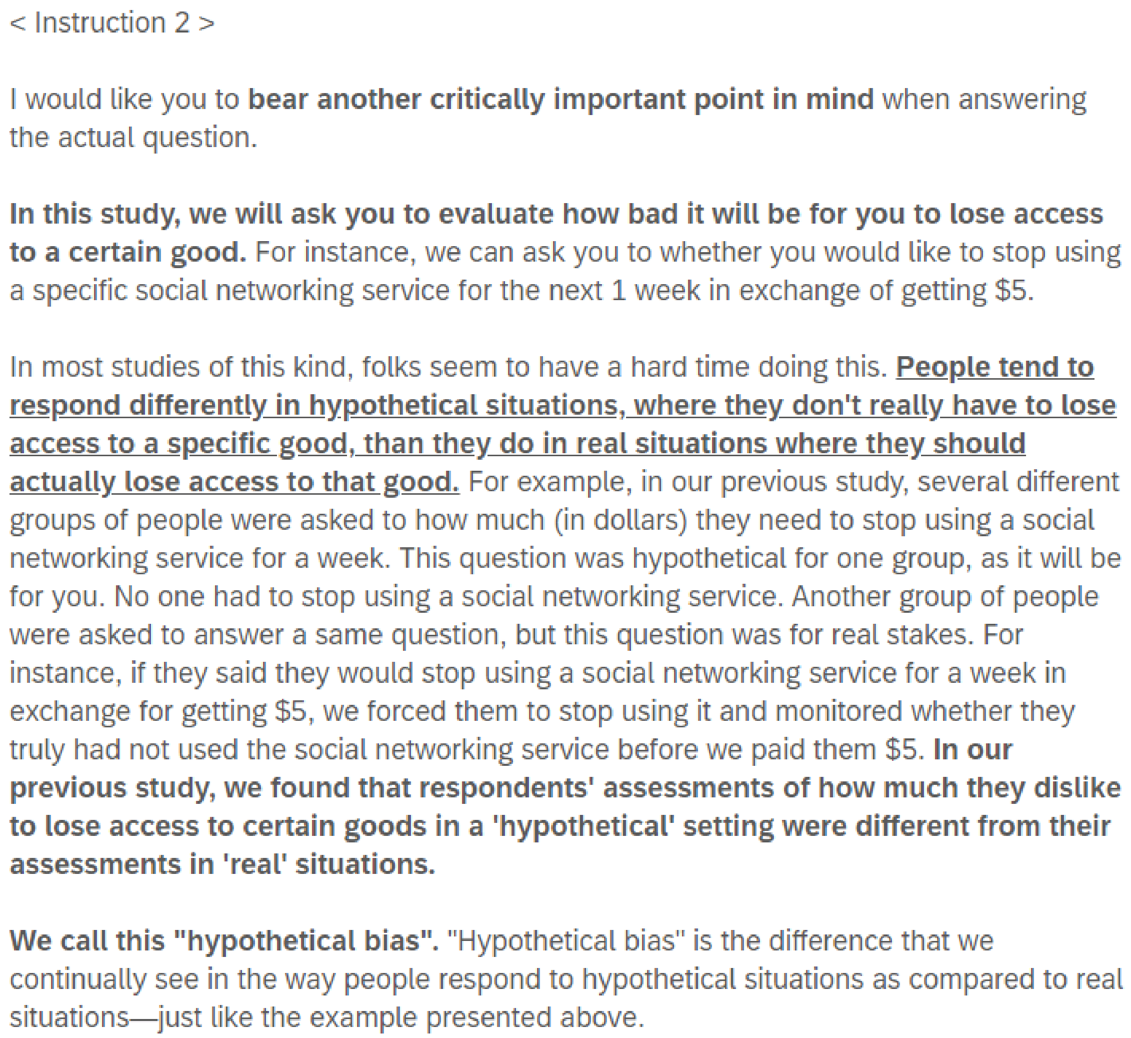
This question is **for real stakes**:

Would you be willing to **avoid using "Facebook" for 1 week in exchange for getting $1**?

○ **YES**

○ **NO**

Figure Appendix 12: Cheap Talk Part in C-BTS in Measuring the WTA Using Binary Discrete Choices



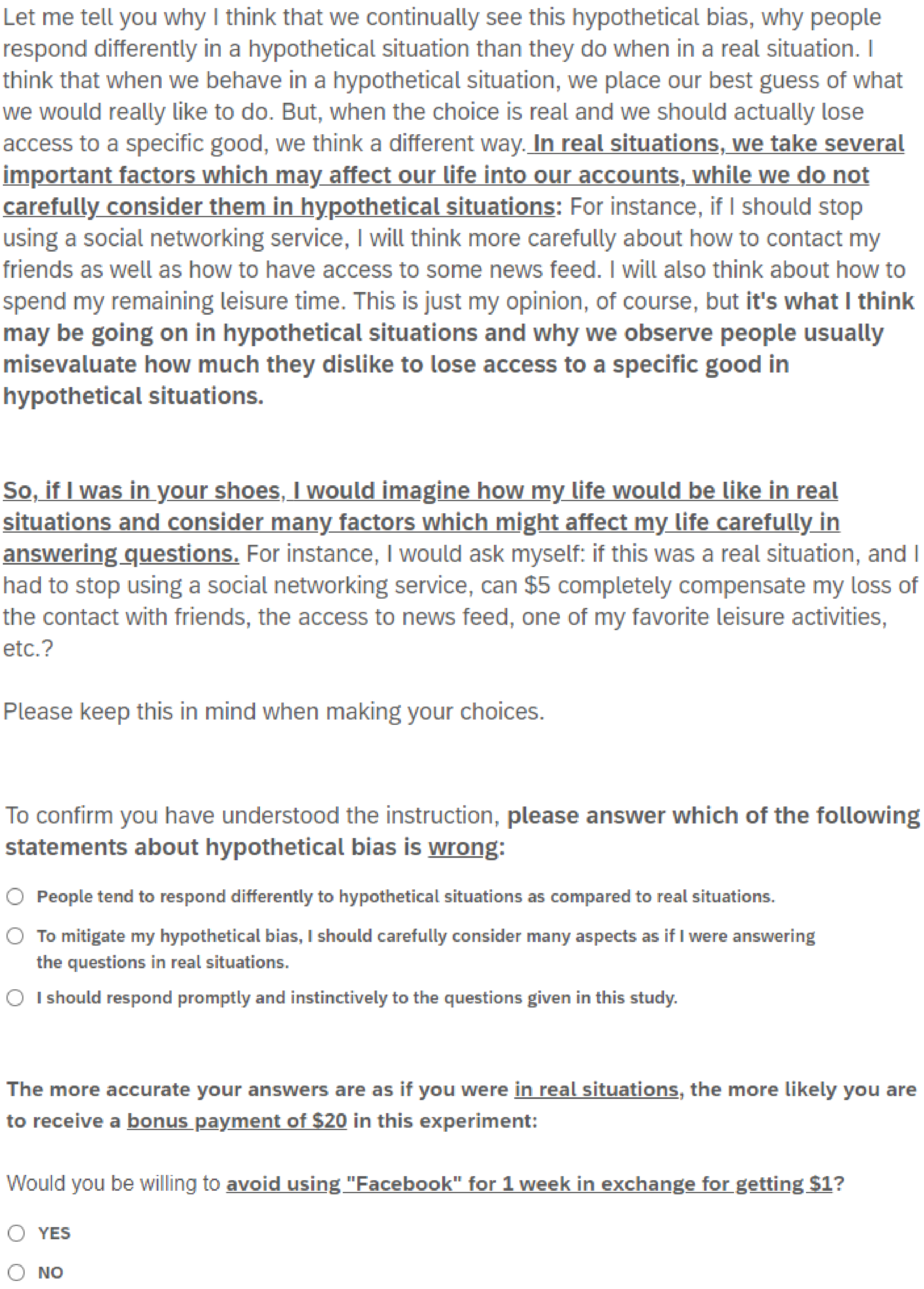
< Instruction 2 >

I would like you to **bear another critically important point in mind** when answering the actual question.

In this study, we will ask you to evaluate how bad it will be for you to lose access to a certain good. For instance, we can ask you to whether you would like to stop using a specific social networking service for the next 1 week in exchange of getting $5.

In most studies of this kind, folks seem to have a hard time doing this. **People tend to respond differently in hypothetical situations, where they don't really have to lose access to a specific good, than they do in real situations where they should actually lose access to that good.** For example, in our previous study, several different groups of people were asked to how much (in dollars) they need to stop using a social networking service for a week. This question was hypothetical for one group, as it will be for you. No one had to stop using a social networking service. Another group of people were asked to answer a same question, but this question was for real stakes. For instance, if they said they would stop using a social networking service for a week in exchange for getting $5, we forced them to stop using it and monitored whether they truly had not used the social networking service before we paid them $5. **In our previous study, we found that respondents' assessments of how much they dislike to lose access to certain goods in a 'hypothetical' setting were different from their assessments in 'real' situations.**

We call this "hypothetical bias". "Hypothetical bias" is the difference that we continually see in the way people respond to hypothetical situations as compared to real situations—just like the example presented above.



Let me tell you why I think that we continually see this hypothetical bias, why people respond differently in a hypothetical situation than they do when in a real situation. I think that when we behave in a hypothetical situation, we place our best guess of what we would really like to do. But, when the choice is real and we should actually lose access to a specific good, we think a different way. **In real situations, we take several important factors which may affect our life into our accounts, while we do not carefully consider them in hypothetical situations:** For instance, if I should stop using a social networking service, I will think more carefully about how to contact my friends as well as how to have access to some news feed. I will also think about how to spend my remaining leisure time. This is just my opinion, of course, but **it's what I think may be going on in hypothetical situations and why we observe people usually misevaluate how much they dislike to lose access to a specific good in hypothetical situations.**

So, if I was in your shoes, I would imagine how my life would be like in real situations and consider many factors which might affect my life carefully in answering questions. For instance, I would ask myself: if this was a real situation, and I had to stop using a social networking service, can $5 completely compensate my loss of the contact with friends, the access to news feed, one of my favorite leisure activities, etc.?

Please keep this in mind when making your choices.

To confirm you have understood the instruction, **please answer which of the following statements about hypothetical bias is wrong:**

○ **People tend to respond differently to hypothetical situations as compared to real situations.**

○ **To mitigate my hypothetical bias, I should carefully consider many aspects as if I were answering the questions in real situations.**

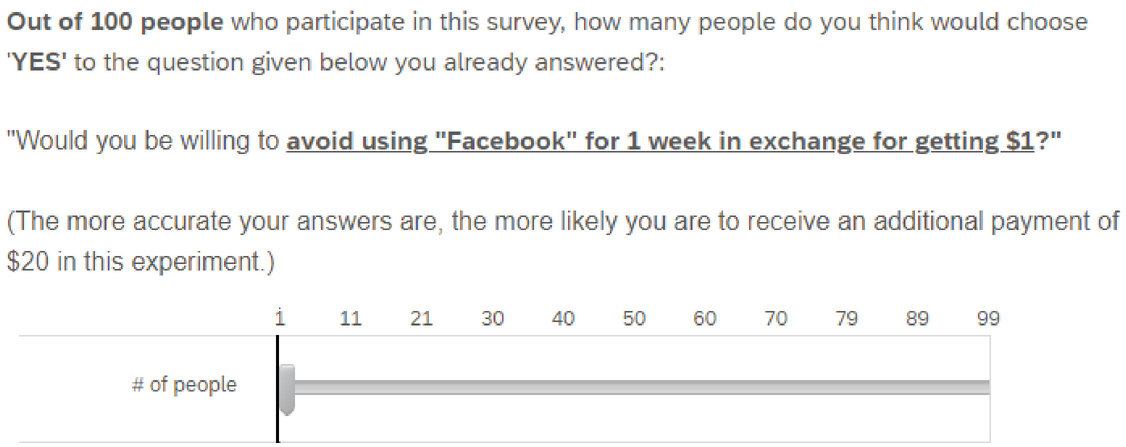
○ **I should respond promptly and instinctively to the questions given in this study.**

The more accurate your answers are as if you were in real situations, the more likely you are to receive a bonus payment of $20 in this experiment:

Would you be willing to **avoid using "Facebook" for 1 week in exchange for getting $1**?

○ **YES**

○ **NO**

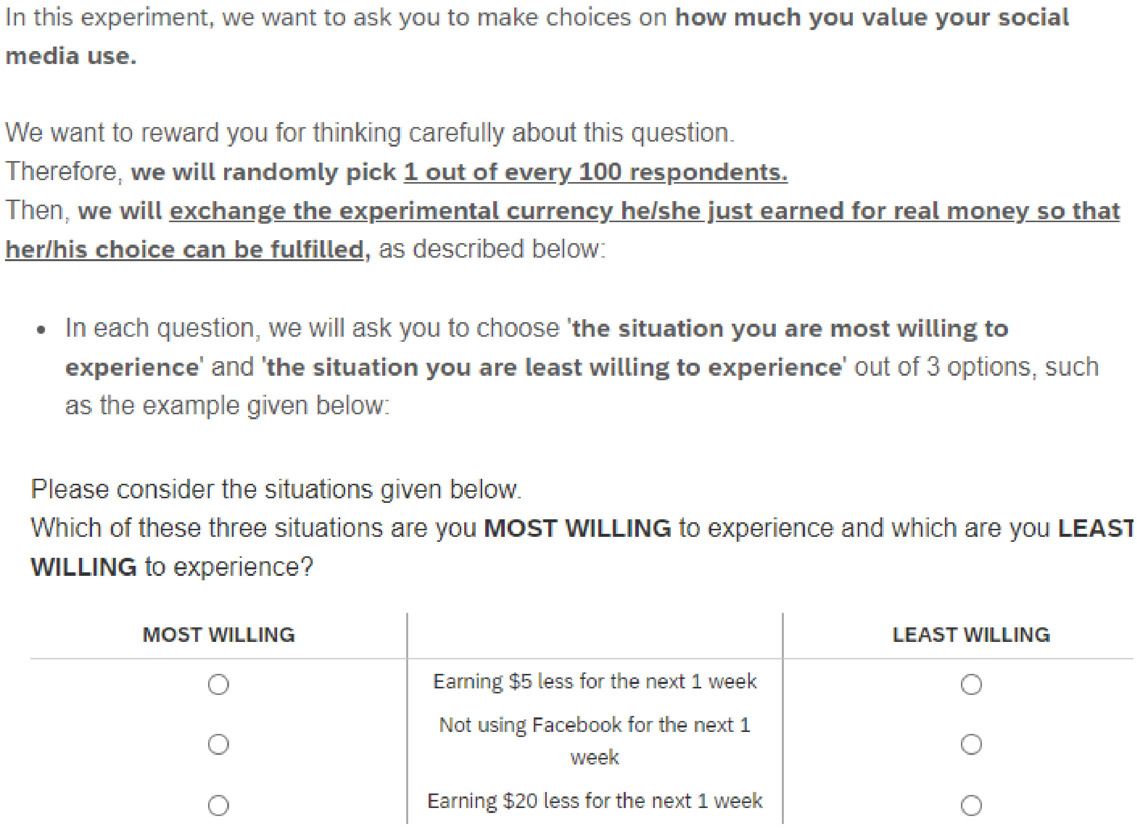


Out of 100 people who participate in this survey, how many people do you think would choose '**YES**' to the question given below you already answered?:

"Would you be willing to **avoid using "Facebook" for 1 week in exchange for getting $1?"**

(The more accurate your answers are, the more likely you are to receive an additional payment of $20 in this experiment.)

Figure Appendix 13: Real Group in Measuring the WTP Using the BWS Format



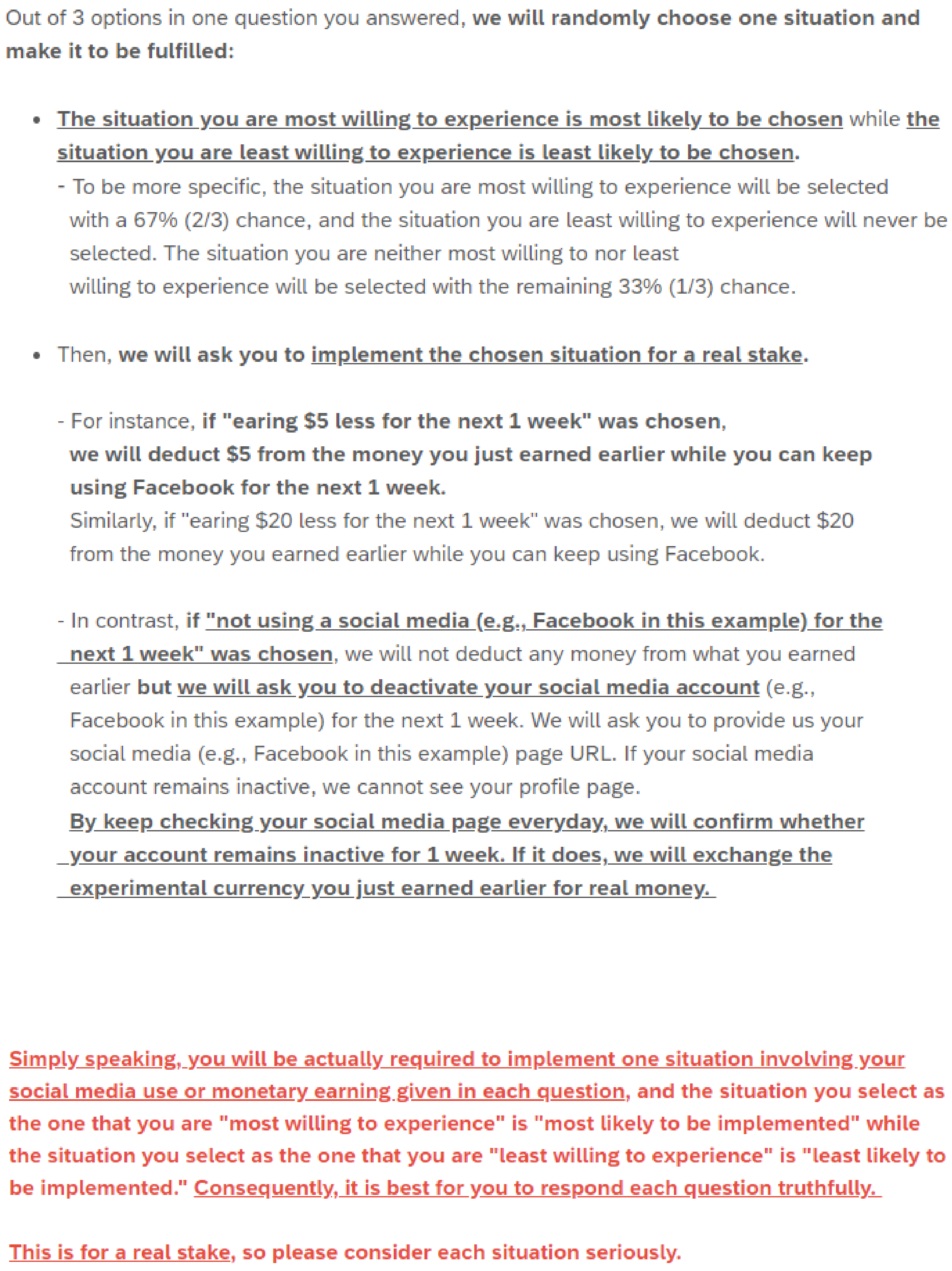
In this experiment, we want to ask you to make choices on **how much you value your social media use.**

We want to reward you for thinking carefully about this question.
Therefore, **we will randomly pick 1 out of every 100 respondents.**
Then, **we will exchange the experimental currency he/she just earned for real money so that her/his choice can be fulfilled,** as described below:

- In each question, we will ask you to choose '**the situation you are most willing to experience**' and '**the situation you are least willing to experience**' out of 3 options, such as the example given below:

Please consider the situations given below.
Which of these three situations are you **MOST WILLING** to experience and which are you **LEAST WILLING** to experience?

MOST WILLING		LEAST WILLING
○	Earning $5 less for the next 1 week	○
○	Not using Facebook for the next 1 week	○
○	Earning $20 less for the next 1 week	○



Out of 3 options in one question you answered, **we will randomly choose one situation and make it to be fulfilled:**

- **The situation you are most willing to experience is most likely to be chosen** while **the situation you are least willing to experience is least likely to be chosen.**
  - To be more specific, the situation you are most willing to experience will be selected with a 67% (2/3) chance, and the situation you are least willing to experience will never be selected. The situation you are neither most willing to nor least willing to experience will be selected with the remaining 33% (1/3) chance.

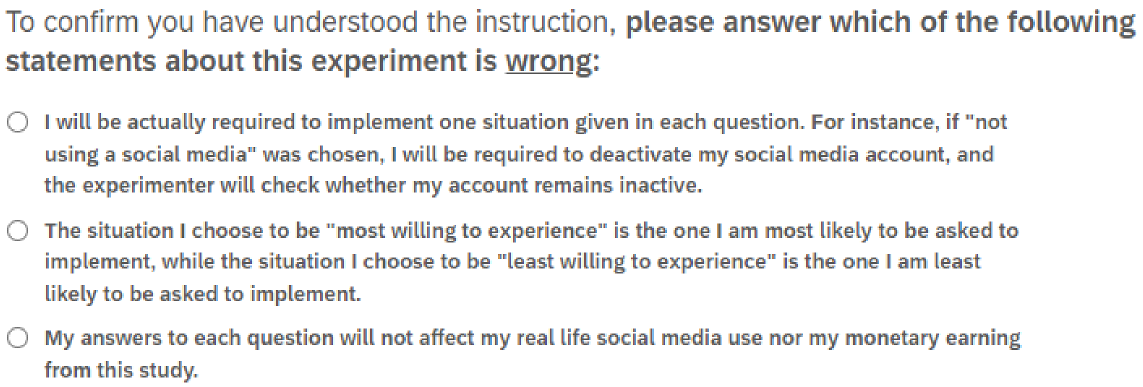
- Then, **we will ask you to implement the chosen situation for a real stake.**

  - For instance, **if "earing $5 less for the next 1 week" was chosen, we will deduct $5 from the money you just earned earlier while you can keep using Facebook for the next 1 week.** Similarly, if "earing $20 less for the next 1 week" was chosen, we will deduct $20 from the money you earned earlier while you can keep using Facebook.

  - In contrast, **if "not using a social media (e.g., Facebook in this example) for the next 1 week" was chosen**, we will not deduct any money from what you earned earlier **but we will ask you to deactivate your social media account** (e.g., Facebook in this example) for the next 1 week. We will ask you to provide us your social media (e.g., Facebook in this example) page URL. If your social media account remains inactive, we cannot see your profile page. **By keep checking your social media page everyday, we will confirm whether your account remains inactive for 1 week. If it does, we will exchange the experimental currency you just earned earlier for real money.**

Simply speaking, you will be actually required to implement one situation involving your social media use or monetary earning given in each question, and the situation you select as the one that you are "most willing to experience" is "most likely to be implemented" while the situation you select as the one that you are "least willing to experience" is "least likely to be implemented." Consequently, it is best for you to respond each question truthfully.

This is for a real stake, so please consider each situation seriously.

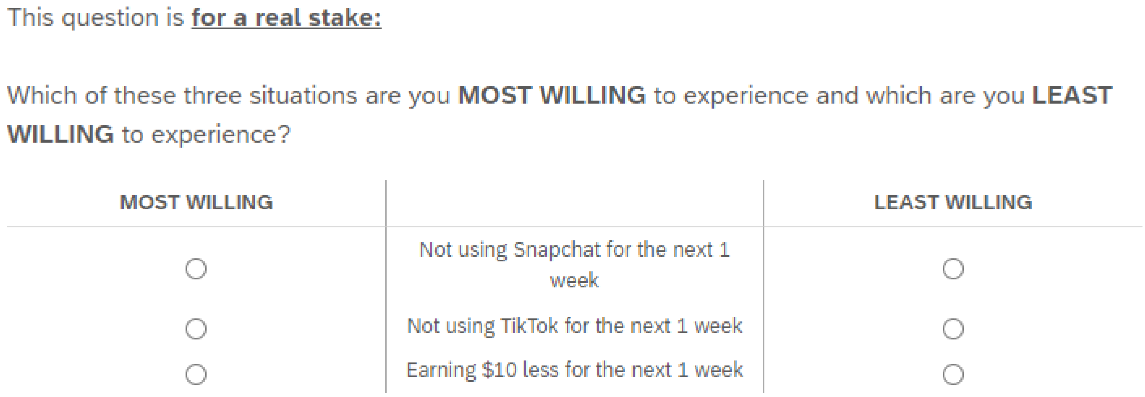


To confirm you have understood the instruction, **please answer which of the following statements about this experiment is wrong:**

○ **I will be actually required to implement one situation given in each question. For instance, if "not using a social media" was chosen, I will be required to deactivate my social media account, and the experimenter will check whether my account remains inactive.**

○ **The situation I choose to be "most willing to experience" is the one I am most likely to be asked to implement, while the situation I choose to be "least willing to experience" is the one I am least likely to be asked to implement.**

○ **My answers to each question will not affect my real life social media use nor my monetary earning from this study.**

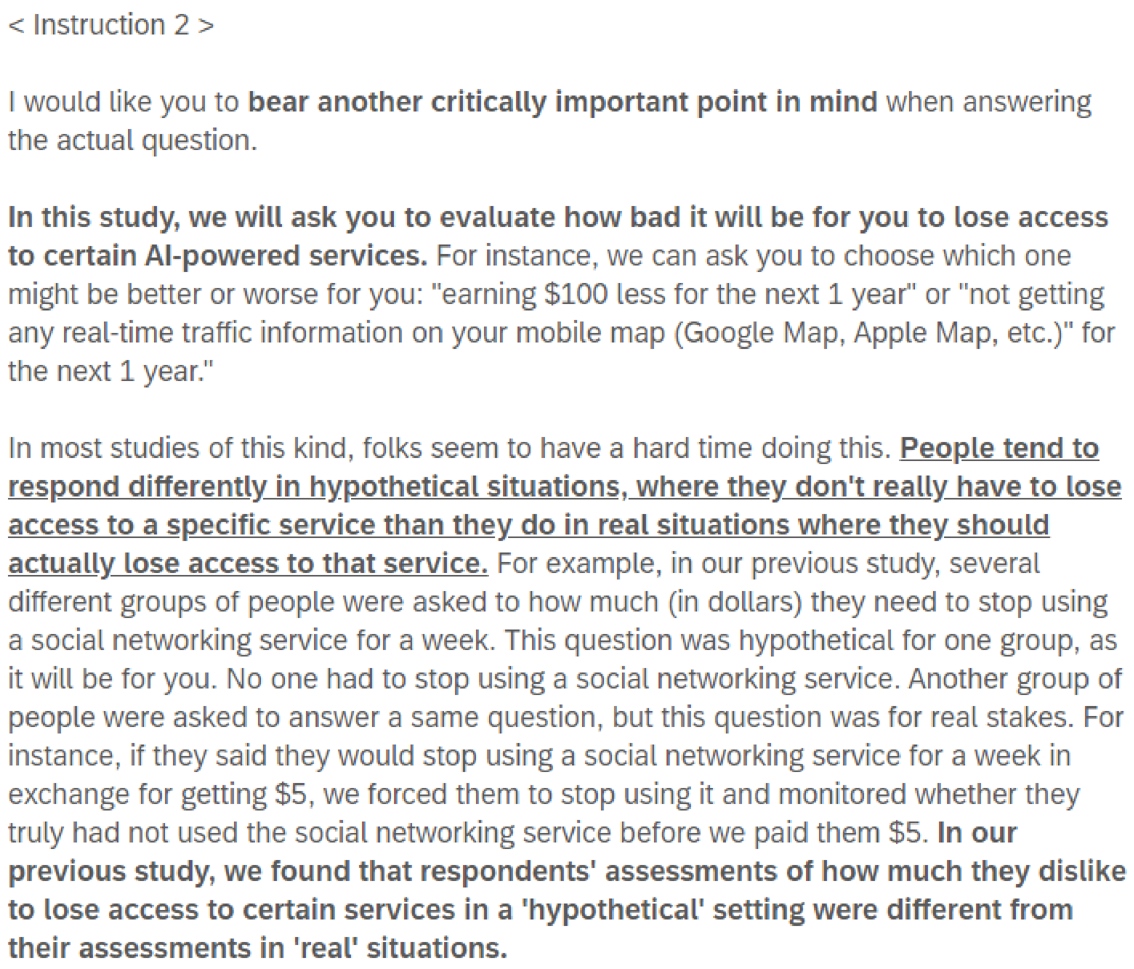


This question is **for a real stake:**

Which of these three situations are you **MOST WILLING** to experience and which are you **LEAST WILLING** to experience?

MOST WILLING		LEAST WILLING
○	Not using Snapchat for the next 1 week	○
○	Not using TikTok for the next 1 week	○
○	Earning $10 less for the next 1 week	○

Figure Appendix 14: Cheap Talk Part in C-BTS in Measuring the Consumer Value of AI-powered Services

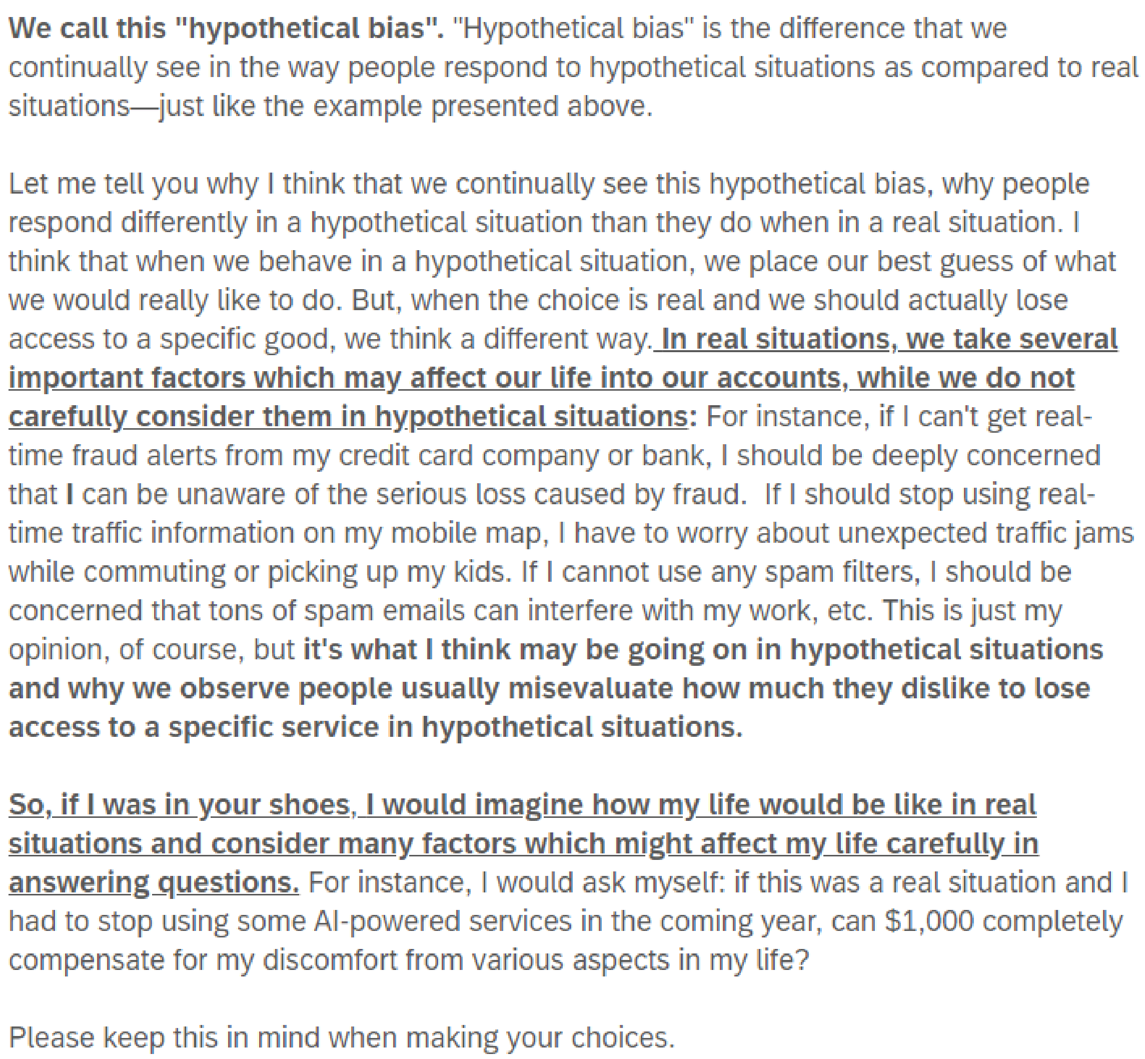


< Instruction 2 >

I would like you to **bear another critically important point in mind** when answering the actual question.

In this study, we will ask you to evaluate how bad it will be for you to lose access to certain AI-powered services. For instance, we can ask you to choose which one might be better or worse for you: "earning $100 less for the next 1 year" or "not getting any real-time traffic information on your mobile map (Google Map, Apple Map, etc.)" for the next 1 year."

In most studies of this kind, folks seem to have a hard time doing this. **People tend to respond differently in hypothetical situations, where they don't really have to lose access to a specific service than they do in real situations where they should actually lose access to that service.** For example, in our previous study, several different groups of people were asked to how much (in dollars) they need to stop using a social networking service for a week. This question was hypothetical for one group, as it will be for you. No one had to stop using a social networking service. Another group of people were asked to answer a same question, but this question was for real stakes. For instance, if they said they would stop using a social networking service for a week in exchange for getting $5, we forced them to stop using it and monitored whether they truly had not used the social networking service before we paid them $5. **In our previous study, we found that respondents' assessments of how much they dislike to lose access to certain services in a 'hypothetical' setting were different from their assessments in 'real' situations.**

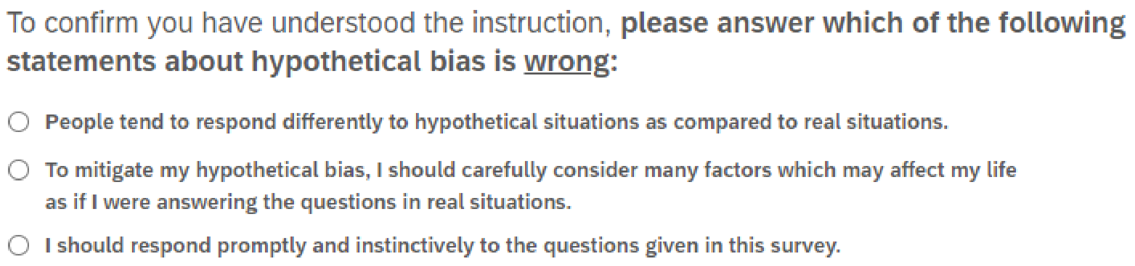


We call this "hypothetical bias". "Hypothetical bias" is the difference that we continually see in the way people respond to hypothetical situations as compared to real situations—just like the example presented above.

Let me tell you why I think that we continually see this hypothetical bias, why people respond differently in a hypothetical situation than they do when in a real situation. I think that when we behave in a hypothetical situation, we place our best guess of what we would really like to do. But, when the choice is real and we should actually lose access to a specific good, we think a different way. **In real situations, we take several important factors which may affect our life into our accounts, while we do not carefully consider them in hypothetical situations:** For instance, if I can't get real-time fraud alerts from my credit card company or bank, I should be deeply concerned that **I** can be unaware of the serious loss caused by fraud. If I should stop using real-time traffic information on my mobile map, I have to worry about unexpected traffic jams while commuting or picking up my kids. If I cannot use any spam filters, I should be concerned that tons of spam emails can interfere with my work, etc. This is just my opinion, of course, but **it's what I think may be going on in hypothetical situations and why we observe people usually misevaluate how much they dislike to lose access to a specific service in hypothetical situations.**

So, if I was in your shoes, I would imagine how my life would be like in real situations and consider many factors which might affect my life carefully in answering questions. For instance, I would ask myself: if this was a real situation and I had to stop using some AI-powered services in the coming year, can $1,000 completely compensate for my discomfort from various aspects in my life?

Please keep this in mind when making your choices.

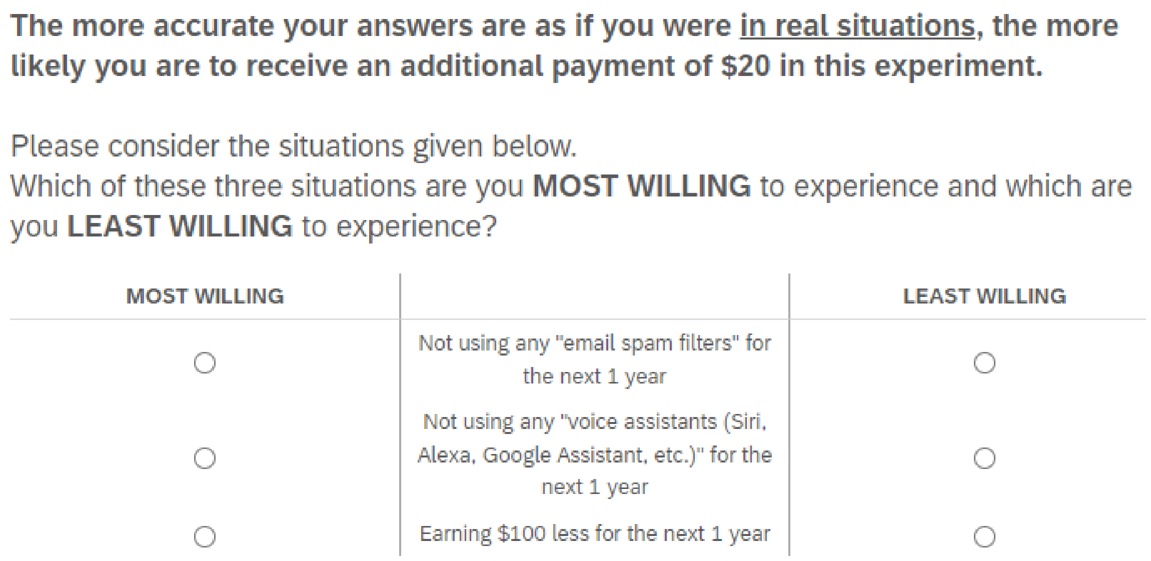


To confirm you have understood the instruction, **please answer which of the following statements about hypothetical bias is wrong:**

○ **People tend to respond differently to hypothetical situations as compared to real situations.**

○ **To mitigate my hypothetical bias, I should carefully consider many factors which may affect my life as if I were answering the questions in real situations.**

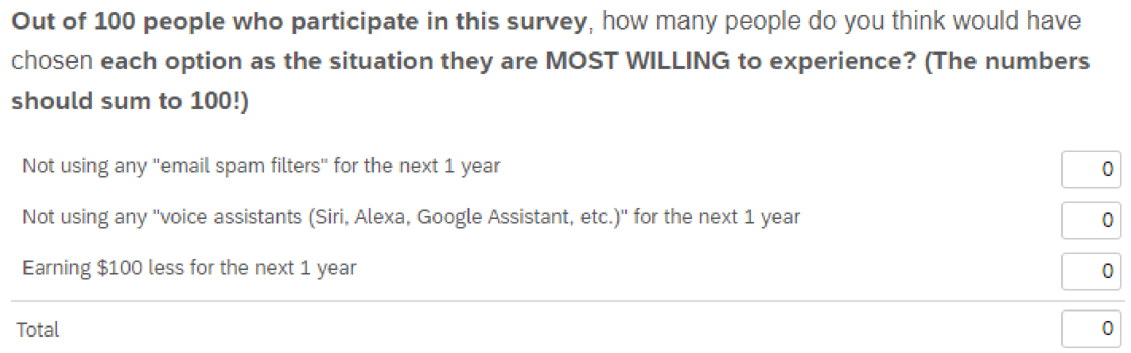
○ **I should respond promptly and instinctively to the questions given in this survey.**



The more accurate your answers are as if you were in real situations, the more likely you are to receive an additional payment of $20 in this experiment.

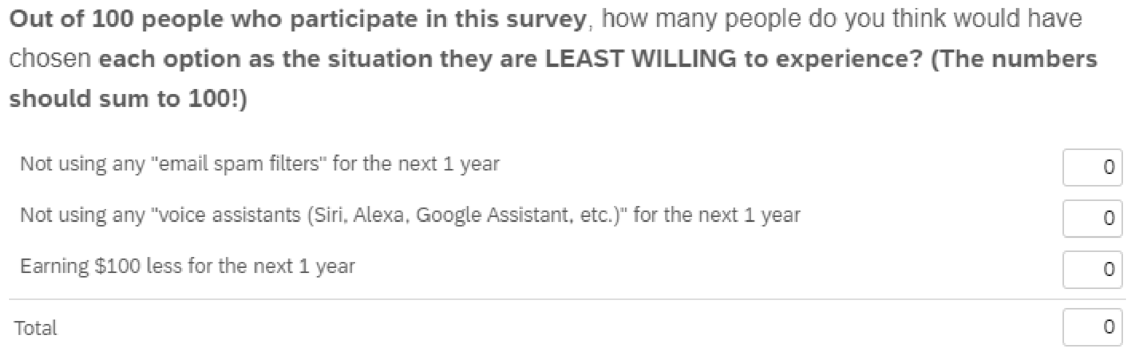
Please consider the situations given below.
Which of these three situations are you **MOST WILLING** to experience and which are you **LEAST WILLING** to experience?

MOST WILLING		LEAST WILLING
○	Not using any "email spam filters" for the next 1 year	○
○	Not using any "voice assistants (Siri, Alexa, Google Assistant, etc.)" for the next 1 year	○
○	Earning $100 less for the next 1 year	○



Out of 100 people who participate in this survey, how many people do you think would have chosen **each option as the situation they are MOST WILLING to experience? (The numbers should sum to 100!)**

Not using any "email spam filters" for the next 1 year	0
Not using any "voice assistants (Siri, Alexa, Google Assistant, etc.)" for the next 1 year	0
Earning $100 less for the next 1 year	0
Total	0



Out of 100 people who participate in this survey, how many people do you think would have chosen **each option as the situation they are LEAST WILLING to experience? (The numbers should sum to 100!)**

Not using any "email spam filters" for the next 1 year	0
Not using any "voice assistants (Siri, Alexa, Google Assistant, etc.)" for the next 1 year	0
Earning $100 less for the next 1 year	0
Total	0