Hypothetical bias in the SG and TTO

AURÉLIEN BAILLON, PHD[1]

HAN BLEICHRODT, PHD[2]

GEORG GRANIC, PHD[3]

7 DECEMBER 2024

[1]Emlyon business school, Lyon, France. Email: baillon@em-lyon.com

[2]Departamento de Fundamentos de Análisis Económico, University of Alicante, Spain. Email: hanbleichrodt@ua.es.

[3]Erasmus School of Economics, Erasmus University Rotterdam. Email: granic@ese.eur.nl

ORCID IDs:

Han Bleichrodt: https://orcid.org/0000-0002-2700-412X

Author contributions: AB, HB, and GG designed the experiments. GG performed the data collection. HB and GG analyzed the data. HB drafted the article. AB and GG wrote revisions. All authors approved the submitted version of the manuscript. All authors had full access to all the data in the study and take responsibility for the integrity of the data and the analysis.

Background

Quality of life measurements are central in health policy and medical decision making. Common methods are the standard gamble (SG) and the time trade-off (TTO). They typically use hypothetical questions. It is unknown whether this leads to a bias.

Methods

We use the Bayesian truth serum (BTS) to incentivize the SG and TTO. The BTS makes it possible to incentivize questions for which the true answer is unknown. We measure quality of life both with and without incentives in two experiments: one lab study with 498 Dutch students and one online study with 1298 members of the US general population. To give incentives their maximal possibility to work, we deliberately introduce default bias in the US sample.

Results

Incentives made no difference in both experiments: SG and TTO valuations were the same with and without incentives. Defaults affected TTO valuations, but not SG valuations. We found the usual SG-TTO gap: SG exceeded TTO.

Limitations

The BTS assumes that respondents have a common prior and use Bayesian updating. Moreover, it is hard to explain why it is in respondents' best interests to answer truthfully.

Conclusions and implications

Incentives have no affect on SG and TTO measurements. Our results support the current practice to use hypothetical questions in quality of life measurement.

Highlights

- Quality of life measurements are central in health policy.
- They usually include no financial incentives and ar evulnerable to hypothetical bias.
- We test whether hypothetical bias affects standard gamble and time trade-off measurements and find no evidnece for it.
- The common practice in quality of life measurement to use hypothetical questions seems valid.

KEY WORDS: quality of life, health policy, hypothetical bias, standard gamble, time trade-off, Bayesian truth serum.

1. Introduction

Quality of life measurements are central in health policy and medical decision making. They help to decide which new treatments are eligible for public funding and to recommend treatments to patients. Common measurement methods are the standard gamble (SG), in which people are asked to trade-off increases in quality of life against mortality risk, and the time trade-off (TTO), in which people are asked to trade off increases in quality of life against decreases in life duration [1].

It is hard to ask SG and TTO questions for real. We could ask the preferences of patients who face these decisions, but often such data do not exist. Therefore, quality of life measurements typically use hypothetical questions.

Economists doubt whether hypothetical questions reveal true preferences as subjects have no incentives to respond truthfully. Opinions diverge on the importance of this hypothetical bias, but the common recommendation is to use incentivized questions when possible. [2]

This paper explores whether hypothetical bias affects quality of life measurements. If people accept, for instance, a larger mortality risk in hypothetical than in incentivized choices, then the SG will overstate the severity of impaired health. This overstatement biases health policy in the direction of improving quality of life at the expense of life extension.

We study hypothetical bias in quality of life measurements. We use Prelec's [3] Bayesian truth serum (BTS) to incentivize the SG and TTO questions. The BTS makes it possible to incentivize questions for which the true answer is unknown. It is in people's best interests to answer truthfully if they are Bayesian decision makers. We explain the BTS in the next Section.

We ran two experiments: one lab study with 459 students from Erasmus University and a large on-line survey with 1298 subjects from the US general population. Both studies gave the same conclusion: there is no hypothetical bias in quality of life measurements. We found no differences between incentivized and non-incentivized questions. This held even when we deliberately introduced a bias using default answers, which tend to have a strong impact on people's preferences [4,5]. Our paper suggests that the

common practice of using nonincentivized quality of life measurements is fine and introduces no bias.

2. Bayesian truth serum

In the BTS, respondents give two answers: their own choice and a prediction of what others would do. Take the standard gamble choice between living in an impaired health state for sure and a risky treatment giving a p% chance of living in full health and a $(1-p)$% chance of dying. Respondents first state whether for a given probability of full health, say 70%, they would choose the treatment.[1] They then predict the proportion of subjects who would choose the treatment for this probability. Their answer to the first question, their *personal score*, gets rewarded if it is *surprisingly common*: if more people than expected opt for it. For example, the treatment option is surprisingly common when 50% of the subjects choose it and the mean predicted proportion is only 40%. Their second answer, their *prediction*, is scored for accuracy.[2]

The BTS assumes that respondents share a common prior, that they update by Bayes' rule and that they believe that the other respondents do so as well. Respondents treat their own preferences as a sample of one favoring their answer. Consequently, they update their prior and expect that the prevalence of their answer will be underestimated by subjects who do not share their preference. Consequently, their answer will be surprisingly common. Prelec [3] showed that answering truthfully is optimal in the BTS.

Empirical evidence shows that the BTS improves truthful reporting. The BTS led to more admitted questionable research practices [6], to less claimed recognition of non-existent concepts [7], and to a reduction in the bias of using defaults in subjective well-being surveys [8].

---

[1] Some studies ask directly for the subject's indifference probability p. The BTS can also be used in this case. We used choices to elicit p and will explain the BTS for these.

[2] The scoring rule is defined as follows. Let x^r and y^r be the personal score and the prediction of subject r. Let $\bar{x}$ be the mean of the x^r and $\bar{y}$ the geometric mean of the y^r. The total score is defined as $x^r log\frac{\bar{x}}{\bar{y}}+\bar{x} log\frac{y^r}{\bar{x}}$. $log\frac{\bar{x}}{\bar{y}}$ is the subjects' information score. It measures how surprisingly common is their answer.

A limitation of the BTS is that it is hard to explain why truth-telling is optimal. The common practice is to tell respondents that the BTS was designed by a professor from MIT, one of the most prestigious universities in the world, and the paper introducing it was published in Science, a leading journal in academia. We also used this explanation, which is sometimes called "intimidation." Baillon [9] and Cvitanich et al. [10] developed modifications of the BTS, which are easier to explain (but have their own limitations).

3. First experiment: the lab study

3.1. Respondents and health states

Respondents were 459 students from Erasmus University. We used three health states, A, B, and full health. They were described by the EuroQol EQ-5D-5L system, which is widely-used in quality of life measurement. The EQ-5D-5L describes health states in terms of five attributes: mobility, self-care, ability to perform usual activities, pain, and anxiety/depression. The attributes can take 5 values with 1 the best and 5 the worst. Health state A corresponds to the state 23133, health state B to 33543, and full health to 11111. To get some feeling for the health states, respondents first valued them on a rating scale with Death at value 0 and full health at 100.

Respondents then answered 3 SG and 3 TTO questions for both health states A and B. In the SG, they chose between living in either impaired health state A or health state B for 13 years for sure (followed by death) and a risky treatment giving a probability p of living in full health for 13 years (followed by death) and a probability $1 - p$ of immediate death. There were 11 SG questions in total for both A and B. For A, p varied from 0.40 to 0.90 in steps of 0.05. For B, from 0.25 to 0.75 in steps of 0.05. We randomly chose the 3 SG questions the respondents answered for each health state from the set of 11 possible questions.

In the TTO, subjects chose between 13 years in A or B (followed by death) and x years in full health. There were 7 TTO questions with x varying between 5 and 11 years in steps of 1 year for A and between 3 and 9 years in steps of 1 year for B. We chose the TTO questions respondents answered randomly from the set of 7 possible questions.

We randomized the order of the SG and the TTO questions, but we did not intersperse SG and TTO questions.

3.2. Experimental treatments

There were 3 experimental treatments. The *Control* treatment (N =158) used the normal SG and TTO. Respondents only stated their preferred option and made no predictions. Respondents dragged their preferred option into an empty box. Figure 3 gives an example.

Respondents in the other two treatments selected their preferred option and predicted how many out of 100 randomly chosen respondents from the experiment would make the same choice as they. Respondents answered the second task by sliding a bar. Figure 5 gives an example. The *Prediction* treatment ($N = 152$) used no incentives, the *Incentives* treatment ($N = 149$) used the BTS. Respondents were randomly assigned to the different treatments. Table A1 in the appendix shows the sample characteristics. There were no differences in age, gender, EU membership, and study discipline between the groups (all p > .31).

3.3. Incentives

We selected one out of every seven respondents to receive €50 for completing the experiment. In the *Control* and *Prediction* treatments each respondent had a $1/7$ chance to get €50. In the *Incentives* treatments, this chance increased with respondents' BTS score. We told respondents beforehand about this procedure.

3.4. Estimation

We used probit regression with respondent as a random effect to test differences in the probability of choosing the impaired health option across treatments. The dependent variable was 1 if a respondent chose the option with impaired health and 0 otherwise. We estimated mean utility by the parametric method of Bleichrodt and Johannesson [11] and computed confidence intervals by the delta method. As a robustness check we

used Kriström's [12] nonparametric method to estimate median utility. Nonparametric and parametric estimates were similar. We estimated the utilities using maximum likelihood estimation with standard errors clustered at the respondent level.

3.5. Results of the first experiment

Figure 1 shows the proportions of subjects choosing the impaired health state for different probabilities of full health (SG) and numbers of years in full health (TTO). The data satisfy two consistency checks. First, the probability of choosing impaired health decreased with the probability of successful treatment in the SG and with the number of years in full health in the TTO (both p<.001). Second, the proportions choosing impaired health was higher for the better health state A than for B (both p<.001).

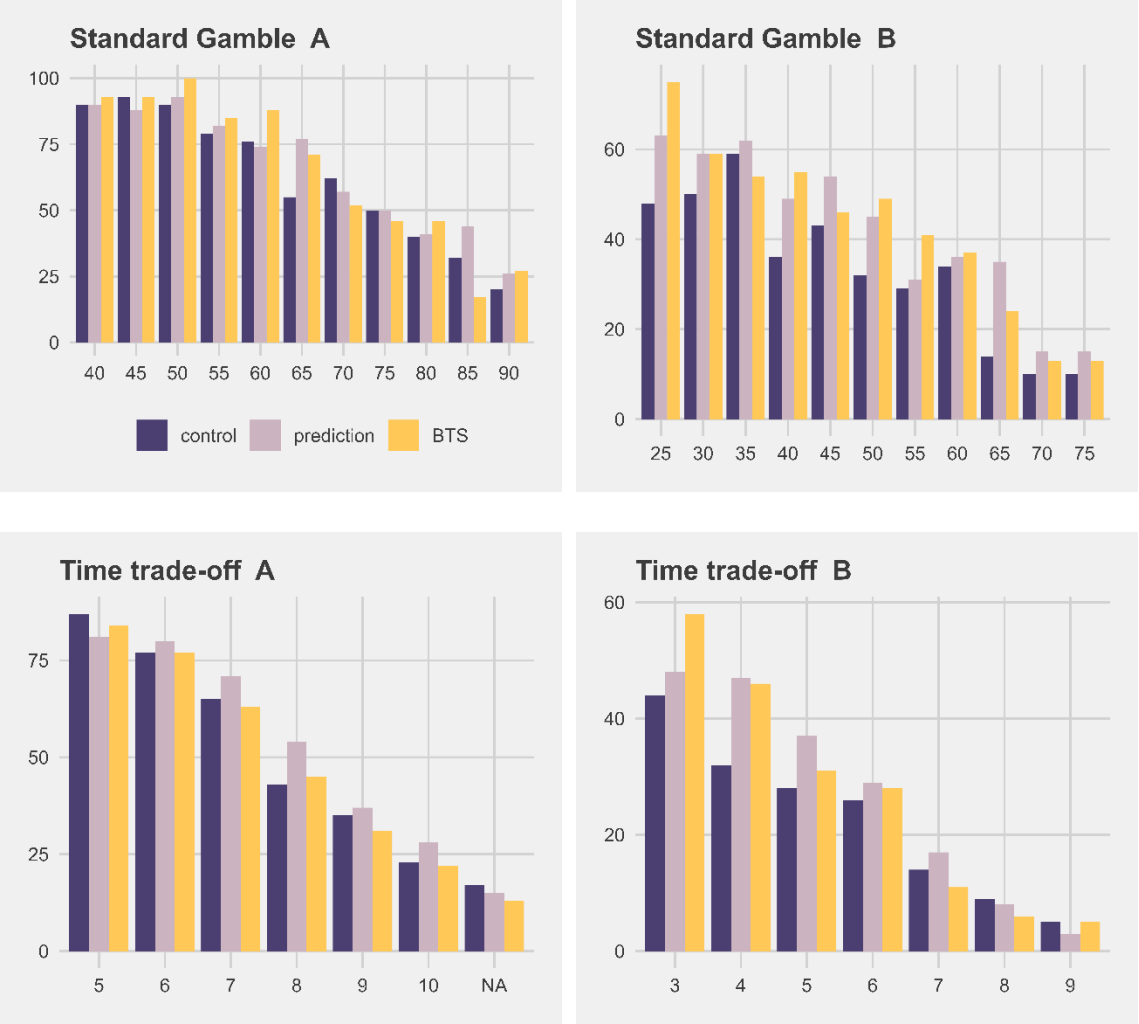


Figure 1: Proportions that choose the impaired health option in the SG and the TTO. Data first experiment. In the SG the x-axis shows the probability of full health. In the TTO, the years in full health.

The proportions choosing impaired health were similar across the three treatments. In the SG, the proportion choosing impaired health was marginally higher in the prediction treatment (p=.061). Incentives had no effect (p=.95). In the TTO, neither the prediction treatment (p=.25) nor incentives (p=.32) had an effect. In the SG, women more likely to choose the impaired health option (p=.017). This is in line with experimental evidence for money that women are more risk averse. There was no gender effect in the TTO (p=.12). Both in the SG (p=.082) and in the TTO (p=.054) there was a marginal age effect: elder respondent were more likely to choose the impaired health option.

Splitting out across the health states, we found no effect of incentives for all health states (all p>.35). For the TTO and for the SG and health state A the prediction task had no impact either (all p>.42) . For the SG and health state B the proportion of respondents choosing B for sure was higher with a prediction task (p=.029). Thinking about what the other respondents would choose made respondents less willing to accept a risk of dying.

For health state A, there was neither an age (both p>.27) nor a gender (both p>.14) effect in the SG and the TTO. For health state B, women were less likely to choose impaired health in the SG (p<.001). There was no gender difference in the TTO (p=.27). In both the SG and the TTO elder subjects were more likely to choose impaired health (p=.059 in the SG and p=.025 in the TTO).

Figure 2 shows the mean utilities of A (left of the dotted line) and B (right of the dotted line) with their 95% confidence interval in the three experimental treatments. Table A1 in the appendix gives full estimation results.

Incentives had no effect on the utilities. For A, we found no difference between the treatments. For B, both the prediction and BTS led to higher utilities than the control. This held for both TTO and SG. This difference was due to including a prediction task. Making a meta-prediction of what other respondents would answer made subjects less

willing to give up life years or accept mortality risk for health state B. Incentivizing the prediction task had no effect.

The confidence intervals were smaller for the BTS. The experimental literature suggests that incentives reduce error. [13] Our data support this.

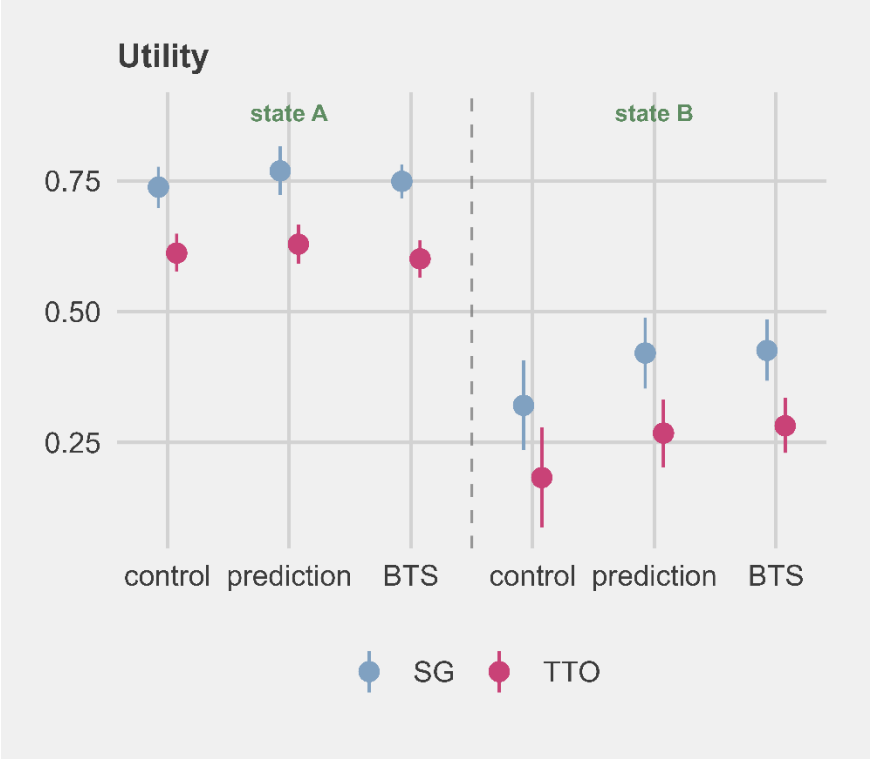


Figure 2: Utilities per treatment. Health state A is on the left of the dotted line, health state B right. First experiment.

Figure 2 shows the common finding that the SG gives higher utilities than the TTO. Incentives did not reduce the SG-TTO gap. The difference between SG and TTO is due to other factors.[14]

4. Experiment 2: the online survey

The first experiment concluded that incentives have no effect on quality of life measurements. The second experiment introduced two changes. First, we used a larger

and more representative sample. Second, we tried to give incentives their best chance to work by introducing a default bias. The common argument favoring incentives is that they lead to more careful reasoning and, consequently, reduce biases. We therefore deliberately introduced a bias by selecting a default option. Empirical evidence suggests that defaults strongly affect preferences [4,5]. We wanted to see whether incentives reduced this bias. If defaults "trap" subjects into choosing a less preferred option, then incentives might counteract this tendency.

4.1. Respondents and health states

We ran an online survey with 1,298 US residents using Prolific. The respondents had a wide variety of socio-demographic backgrounds, but were not fully representative of the US population. Table A3 in the Appendix summarizes the characteristics of the respondents. The survey was part of a larger study testing the impact of incentives on survey data (see Anonymous [8] for details).

The survey used only health state B and full health. B was called A in the experiment, but for ease of reading we call it B to clarify that it is the same health state as B in the first experiment.

In the SG questions, the duration of B and full health was the rest of life. In the TTO questions the duration of B was 10 years (followed by death). There were 6 SG and 6 TTO questions in total. In the SG, the probability of full health p varied from 0.15 to 0.90 in steps of 0.15. In the TTO, the duration in full health varied between 1.5 and 9 years in steps of 1.5 years. Each subject answered 3 randomly selected SG questions and 3 randomly selected TTO questions. We reduced the number of questions compared with the first experiment, because Prolific imposed a time restriction on the duration of the experiment.

4.2. Treatments

In addition to the *Control* ($N = 225$), *Prediction* ($N = 211$), and *Incentives* treatment ($N = 224$), which were the same as in the first experiment, the survey also had these three treatments with a preselected default. In the SG the default was B for sure the rest

of your life and in the TTO it was 10 years in B (followed by death). The *Default* treatment ($N = 209$) was equal to the *Control* treatment with a default, the *Prediction+Default* treatment ($N = 210$) was equal to the *Prediction* treatment with a default, and the *Incentives+Default* treatment ($N = 219$) was equal to the *Incentives* treatment with a default. Table 2 summarizes the different treatments.

As in experiment 1, respondents dragged their preferred option into a box. In the treatments without a default this box was empty. Figure 3 gives an example. In the treatments with a default, the default option was put in the preferred option box. Figure 4 gives an example for the SG. To change their preferences in the default treatments, respondents had to remove the default option from the box and put their preferred option in it.

The prediction task was the same as in the first experiment. Figure 5 gives an example.

Table 2: Treatments in the second experiment.

Treatment	Prediction task	Defaults	BTS incentives
Control	No	No	No
Defaults	No	Yes	No
Prediction	Yes	No	No
Prediction+Defaults	Yes	Yes	No
Incentives	Yes	No	Yes
Incentives+Defaults	Yes	Yes	Yes

Note: the first experiment only included the Control, Prediction, and Incentives treatments.

4.3. Incentives

Respondents in the non-incentivized treatments received a flat fee of $2.58. This was twice the rate that Prolific recommended as fair. The BTS treatments paid respondents by their score. To have the same expected payoff in the three treatments, we restricted BTS payoffs to the interval [$1.29,$3.87]. Respondents knew that their payment

depended on their own BTS score and the score of a random group of 7 other respondents.

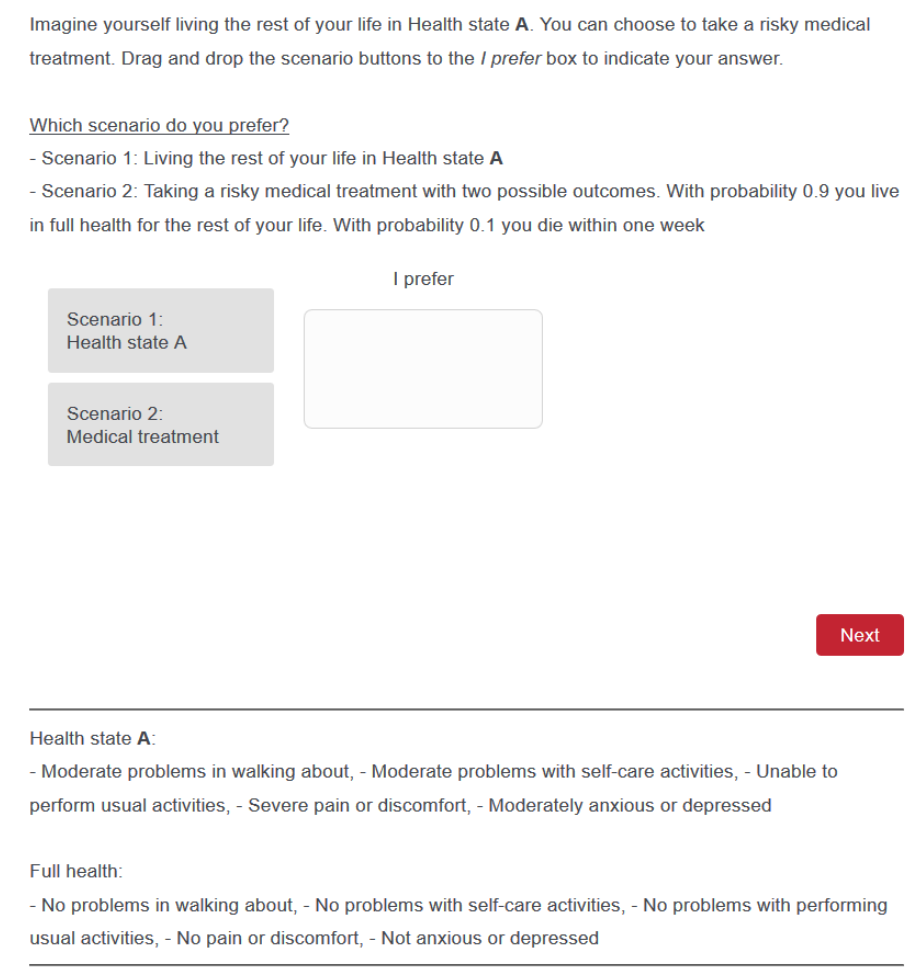


Figure 3: A SG question without defaults.

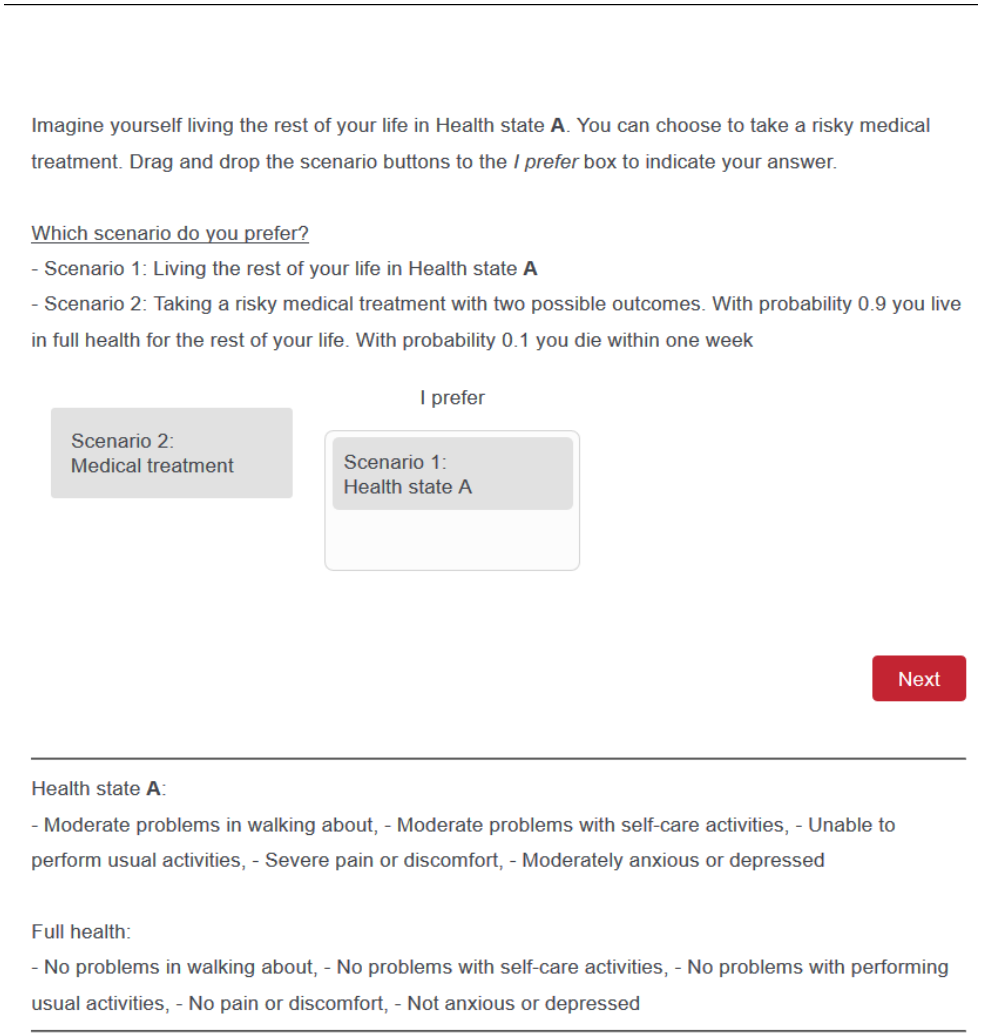


Figure 4: A SG question with defaults.

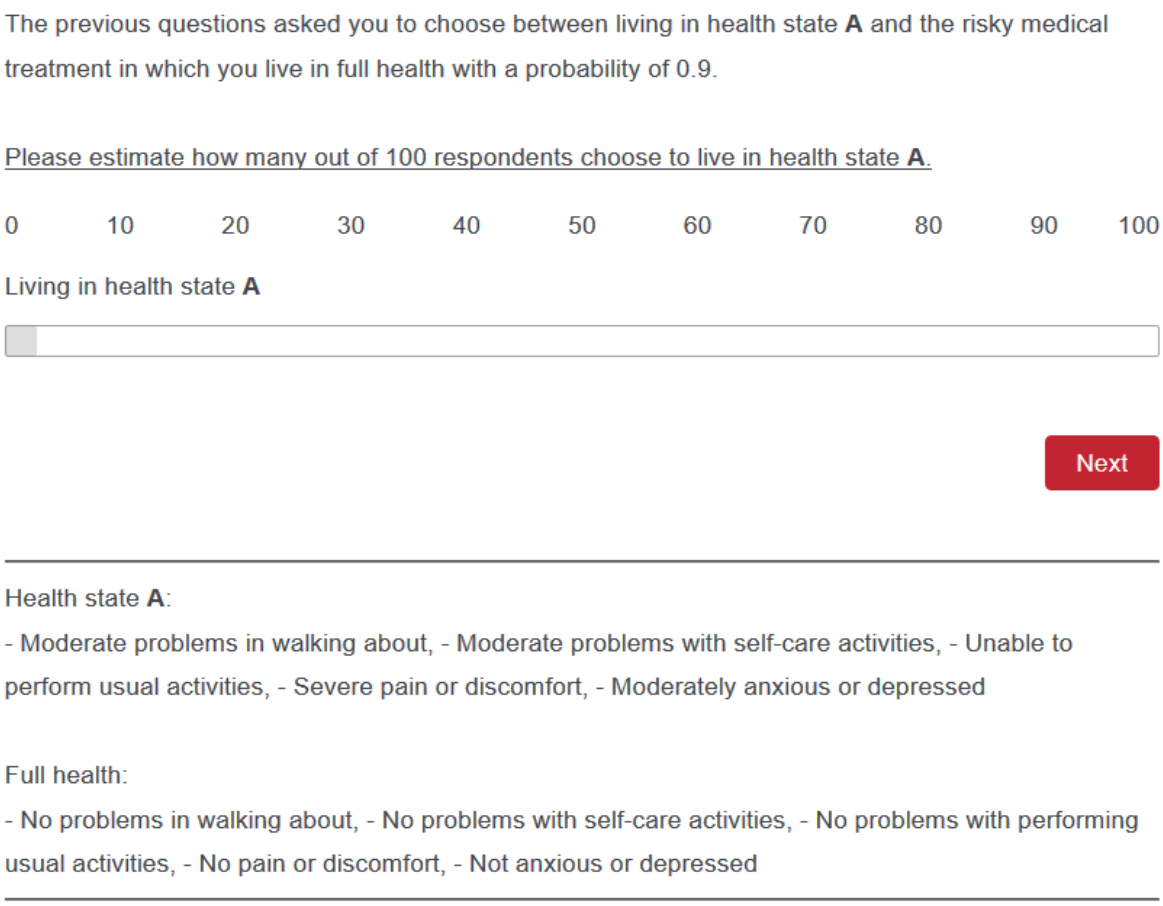


Figure 5: A prediction task for the SG

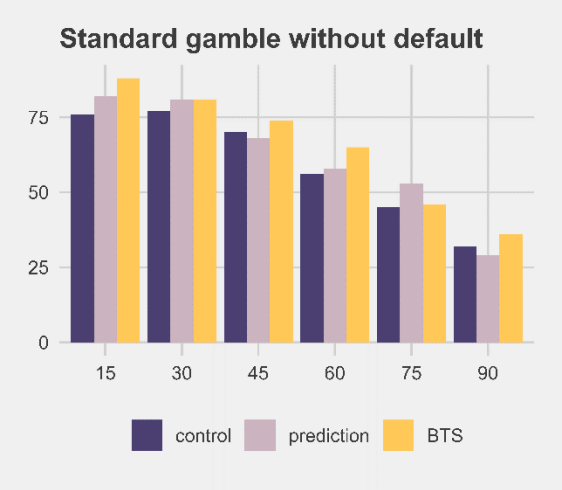
4.4 Results of the second experiment

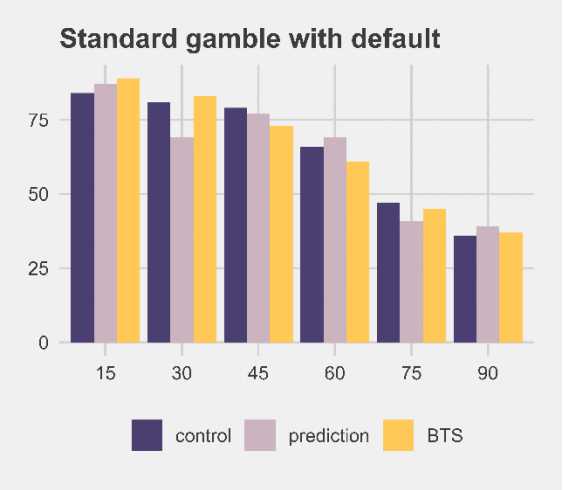
Figure 6 shows the proportions of subjects choosing the impaired health option for various probability of successful treatment (SG) and number of years in full health (TTO). Again, the proportion choosing impaired health decreases with the probability of full health (SG) and the number of years in full health (both p<.001).

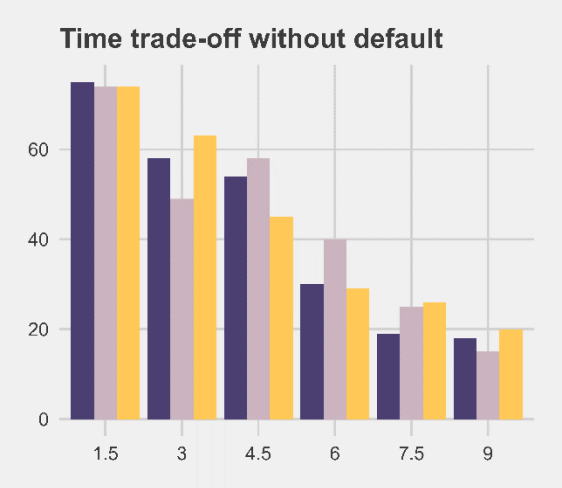
There were no differences between the three treatments. Unlike in the first experiment, including a prediction task had no effect (p=.60 in the SG, p=.65 in the TTO). Like in the first experiment, incentives had no effect (p=.52 in the SG, p=.83 in the TTO). Both in the SG (p=.002) and the TTO (p=.016) older respondents were more likely to choose impaired health. In the TTO, this was also true for women (p<.001). Unlike in the first

experiment, in the SG gender had no effect on the probability of choosing impaired health (p=.13). So we found no evidence that women are more risk averse for health in the second experiment.

Defaults had no effect in the SG. The probability of choosing the default option, impaired health, was the same with and without defaults (p=.13). In the TTO, respondents were more likely to choose the default impaired health (p<.001). We do not know why defaults affected the TTO, but not the SG. It was not caused by respondents spending more time thinking: we could not reject the null that decision time was the same in the SG and the TTO (p=.68). A Bayesian analysis showed very strong support for the null that it was the same (Bayes factor = .02).







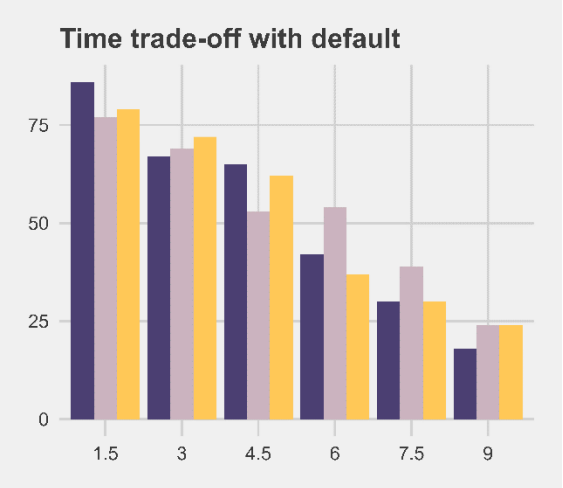


Figure 6: Proportions choosing the impaired health option in the SG and the TTO with and without defaults. Data second experiment. In the SG the x-axis shows the probability of full health. In the TTO, it shows the years in full health.

Figure 7 shows the SG and TTO utilities for the 6 treatments. Again, incentives had no effect on the SG and TTO. Utilities were close and confidence intervals overlapped. The width of the confidence intervals was similar across the treatments suggesting, unlike in the first experiment, that incentives did not reduce noise. The difference between SG and TTO remained and incentives did not reduce it.

The effect of defaults varied across the SG and the TTO. For the SG, defaults had no effect. On the other hand, defaults affected the TTO. Defaults reduced the difference between SG and TTO, because they affected the TTO but not the SG. This is a consequence of making A the default. If we would have made full health the default then defaults would probably have increased the difference.

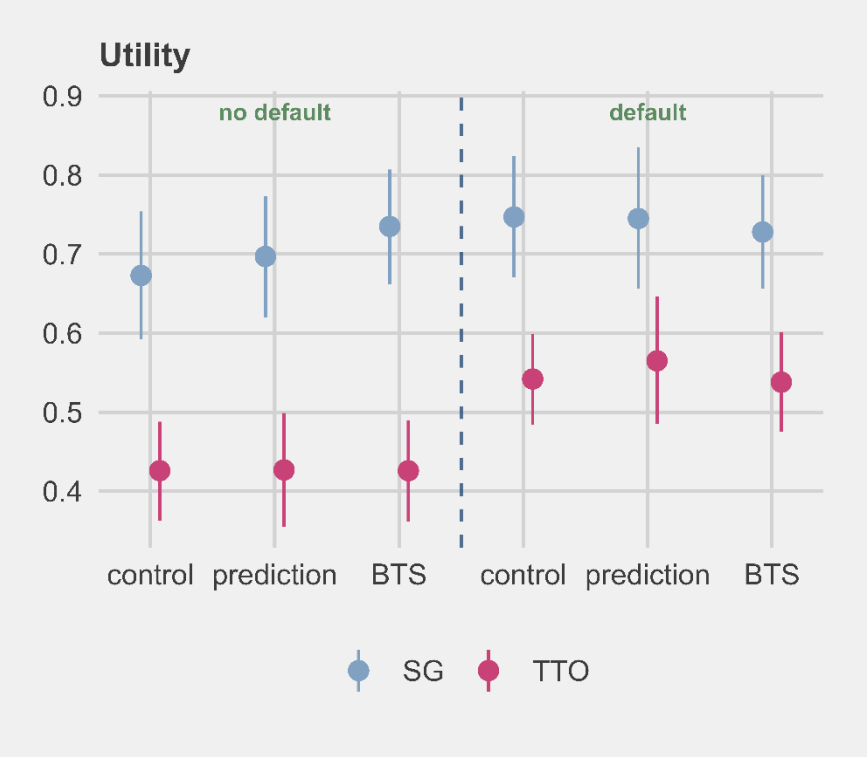


Figure 7: Utility in the second experiment.

5. Conclusion

In two experiments, we found no evidence that incentives affect the SG and TTO. This was even true when we introduced a (default) bias. Incentives did not reduce this bias. Including a prediction task had some effect in some tests, but its effect seemed limited. Our results suggest that people's intrinsic motivation to answer questions involving health is strong enough to make incentives or other extensions unnecessary. We do not claim that incentives never work, but for health they seem to offer no additional benefits.

Future research may wish to address whether our results can be replicated using other incentivized mechanisms than the BTS. The BTS assumes that decision makers share a common prior and update similarly. Moreover, it is hard to explain to respondents why it is in their best interest to respond truthfully. Recent variants relax some of these assumptions (while introducing others) and make it (slightly) easier to explain why responding truthfully is in respondents' best interests.

Overall our results are reassuring for quality of life measurements. They play a central role in health policy. Our study supports its common practice of using nonincentivized questions and suggests that they are not affected by hypothetical bias.

STATEMENTS AND DISCLOSURES

Ethical considerations: Ethical approval was given by the Ethics Committee of the ESE Lab. All respondents gave informed consent before the experiments started.

Funding: Financial support for this study was provided in part by a grant from [insert name(s) of the funding source(s), whether a company, government agency, philanthropic foundation, institute, etc.]. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

Declaration of conflicting interests: The Author(s) declare(s) no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability: The data and a file with the analystics code will be made available upon acceptance.

References

1. Drummond, Mike F., Sculpher, Mark J., Claxton, Karl, Stoddart, Greg L., & Torrance, George W. (2015). *Methods for the economic evaluation of health care programmes*. Oxford university press.

2. Moffatt, Peter, Starmer, Chris, Sugden, Robert, Bardsley, Nick, Cubitt, Robin, & Loomes, Graham (2020). *Experimental economics: Rethinking the rules*. Princeton University Press.
3. Prelec, Drazen (2004). A Bayesian truth serum for subjective data. *Science*, *306*(5695), 462-466.
4. Johnson, Eric J., & Goldstein, Daniel (2003). Do defaults save lives? *Science*, *302*(5649), 1338-1339.
5. Jachimowicz, Jon M., Duncan, Shannon, Weber, Elke U., & Johnson, Eric J. (2019). When and why defaults influence decisions: A meta-analysis of default effects. *Behavioural Public Policy*, *3*(2), 159-186.
6. John, Leslie K., Loewenstein, George, & Prelec, Drazen (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524-532.
7. Weaver, Ray, & Prelec, Drazen (2013). Creating truth-telling incentives with the Bayesian truth serum. *Journal of Marketing Research*, *50*(3), 289-302.
8. Anonymous (2022).
9. Baillon, Aurélien (2017). Bayesian markets to elicit private information. *Proceedings of the National Academy of Sciences*, *114*(30), 7958-7962.
10. Cvitanić, Jakša, Prelec, Drazen, Riley, Blake, & Tereick, Benjamin (2019). Honesty via choice-matching. *American Economic Review: Insights*, *1*(2), 179-92.
11. Bleichrodt, Han, & Johannesson, Magnus (2001). Time preference for health: A test of stationarity versus decreasing timing aversion. *Journal of Mathematical Psychology*, *45*(2), 265-282.
12. Kriström, Bengt (1990). A non-parametric approach to the estimation of welfare measures in discrete response valuation studies. *Land Economics*, *66*(2), 135-139.
13. Camerer, Colin F., & Hogarth, Robin M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, *19*(1), 7-42.
14. Bleichrodt, Han (2002). A new explanation for the difference between time trade-off utilities and standard gamble utilities. *Health Economics*, *11*(5), 447-456.

Appendix

	Control	Prediction	BTS
Number	158	152	149
Male	82	77	77
Proportion male	52%	51%	52%
Mean age in years	21.9	22.0	22.0
Proportion of EU citizens	87%	83%	82%
Proportion with economics background	82%	85%	83%

Table A1: respondents in the first experiment

Table A1 shows the statistics of the respondents in the first experiment. There were no differences across treatments in age, gender, EU citizenship, economics background (all *p-values > 0.31*).

Table A2 shows the health utility estimates in the first experiment. **.** The dependent variable is 1 if the respondent chose the impaired health option. We used Bleichrodt and Johannesson's [11] method to find the parametric estimates and Kriström's [12] to find the nonparametric estimates. Parametric differences in utility across SG and TTO come from bootstrapping.

Experimental treatment	Control	Prediction	Incentives	Control	Prediction	Incentives
Health State	A	A	A	B	B	B
Probit coefficients estimates (with cluster robust std. err.)						
Dep. Var. in all models: dummy, equal to 1 if respondent rejects medical treatment						
Standard Gamble						
Probability successful treatment	-4.570 ***	-4.089 ***	-5.452 ***	-2.693 ***	-2.755 ***	-3.073 ***
	('0.519)	(0.538)	(0.596)	(0.469)	(0.453)	(0.437)
Intercept	3.371 ***	3.146 ***	4.085 ***	0.866 ***	1.160 ***	1.310 ***
	(0.363)	(0.380)	(0.431)	(0.246)	(0.247)	(0.223)
Time Trade-Off						
Years spent in full health	-0.357 ***	-0.344 ***	-0.370 ***	-0.231 ***	-0.285 ***	-0.328 ***
	(0.038)	(0.040)	(0.040)	(0.035)	(0.035)	(0.038)
Intercept	2.839 ***	2.809 ***	2.888 ***	0.549 **	0.993 ***	1.205 ***
	(0.305)	(0.343)	(0.338)	(0.217)	(0.222)	(0.222)
Population utility levels						
Standard Gamble						
Parametric	0.738	0.769	0.749	0.321	0.421	0.426
95% CI	[0.698, 0.777]	[0.723, 816]	[0.717, 0.781]	[0.236, 0.407]	[0.353, 0.489]	[0.368, 0.485]
Non-parametric median	0.750	0.750	0.720	0.358	0.460	0.432
Time Trade-Off						
Parametric	0.612	0.629	0.601	0.183	0.268	0.282
95% CI	[0.576, 0.649]	[0.592, 0.666]	[0.565, 0.637]	[0.087, 0.279]	[0.203, 0.332]	[0.230, 0.335]
Non-parametric median	0.591	0.633	0.593	0.206	0.221	0.282
Difference in utility levels: SG - TTO						
Parametric (bootstrap)	0.125	0.140	0.148	0.138	0.153	0.144
95% CI	[0.089, 0.159]	[0.104, 0.180]	[0.116, 0.189]	[0.047, 0.249]	[0.093, 0.216]	[0.089, 0.199]
Non-parametric median	0.159	0.117	0.127	0.152	0.239	0.150

Table A2: health utility estimates in the first experiment

*** significant at 0.1% level, ** significant at 1% level.

Table A3 shows the statistics of the respondents in the second experiment. The treatment groups were similar in age, gender, and income (all $p > 0.44$).

Variable	Control	Prediction	Incentives	Defaults	Defaults+ Prediction	Defaults+ Incentives	Experiment
Number of observation	225	211	224	209	210	219	1298
Mean age in years (std. dev.)	37.3 (11.4)	36.8 (11.9)	37.2 (10.7)	36.5 (10.5)	36.0 (10.6)	36.3 (10.5)	36.7 (11.0)
Gender:							
Female	48.0%	49.3%	51.3%	48.8%	51.4%	53.0%	50.3%
Male	49.8%	48.8%	46.0%	49.3%	47.6%	46.1%	47.9%
Other	2.2%	1.4%	2.2%	1.4%	1.0%	0.5%	1.5%
Not disclosed	0.0%	0.5%	0.4%	0.5%	0.0%	0.5%	0.3%
Income:							
Less than $25,000	22.2%	21.3%	18.8%	21.5%	14.8%	13.2%	18.6%
$25,000 to $49,999	26.7%	20.4%	22.3%	20.1%	24.3%	23.7%	23.0%
$50,000 to $64,999	12.4%	9.5%	12.9%	12.9%	12.9%	11.9%	12.1%
$65,000 to $104,999	16.0%	17.5%	23.7%	25.4%	21.9%	26.0%	21.7%
$105,000 or more	17.8%	26.5%	20.1%	19.1%	23.8%	21.9%	21.5%
Not disclosed	4.9%	4.7%	2.2%	1.0%	2.4%	3.2%	3.1%
Mean time taken in sec (std. dev.)	466 (312)	708 (301)	730 (303)	448 (271)	704 (338)	769 (368)	638 (342)
Mean earnings in $ (std. dev.)	2.58	2.58	2.58 (0.25)	2.58	2.58	2.58 (0.33)	2.58 (0.17)

Table A3: respondents in the second experiment

Table A4 shows the health utility estimates for the second experiment. Again, the dependent variable is 1 if a respondent chose the impaired health option. Parametric estimates come from the method of Bleichrodt and Johannesson [11] , nonparametric from Kriström [12]. We used bootstrapping to find the parametric differences in utility between SG and TTO.

	Control	Prediction	Incentives	Defaults	Prediction + Defaults	Incentives + Defaults
Probit coefficients estimates (with cluster robust std. err.)						
Dep. Var. in all models: dummy, equal to 1 if respondent rejects medical treatment						
Standard Gamble						
Probability successful treatment	-1.725 ***	-1.955 ***	-2.091 ***	-2.001 ***	-1.764 ***	-2.160 ***
	(0.211)	(0.231)	(0.221)	(0.236)	(0.212)	(0.223)
Intercept	1.161 ***	1.362 ***	1.537 ***	1.499 ***	1.315 ***	1.572 ***
	(0.140)	(0.153)	(0.148)	(0.162)	(0.143)	(0.149)
Time Trade-Off						
Years spent in full health	-0.226 ***	-0.198 ***	-0.210 ***	-0.251 ***	-0.180 ***	-0.223 ***
	(0.023)	(0.022)	(0.022)	(0.024)	(0.021)	(0.023)
Intercept	0.961 ***	0.847 ***	0.894 ***	1.360 ***	1.016 ***	1.199 ***
	(0.132)	(0.132)	(0.123)	(0.147)	(0.132)	(0.140)
Population utility levels						
Standard Gamble						
Parametric	0.673	0.697	0.735	0.747	0.745	0.728
95% CI	[0.592, 0.754]	[0.620, 0.773]	[0.662, 0.807]	[0.671, 0.824]	[0.656, 0.835]	[0.656, 0.800]
Non-parametric median	0.681	0.767	0.716	0.725	0.703	0.703
Time Trade-Off						
Parametric	0.426	0.427	0.426	0.542	0.565	0.538
95% CI	[0.363, 0.488]	[0.355, 0.499]	[0.362, 0.490]	[0.484, 0.599]	[0.485, 0.646]	[0.475, 0.601]
Non-parametric median	0.474	0.489	0.408	0.550	0.636	0.523
Difference in utility levels: SG - TTO						
Parametric (bootstrap)	0.247	0.269	0.309	0.206	0.180	0.190
95% CI	[0.184, 0.325]	[0.205, 0.366]	[0.243, 0.381]	[0.139, 0.280]	[0.110, 0.267]	[0.128, 0.259]
Non-parametric median	0.207	0.278	0.308	0.175	0.067	0.180

Table A4: health utility estimates in the second experiment.

*** significant at 0.1% level.