Incentives in Surveys

Aurelien Baillon, Han Bleichrodt, Georg D. Granic
Erasmus University Rotterdam[1]

January 2022

Abstract

Surveys typically use hypothetical questions to measure subjective and unverifiable concepts like happiness and quality of life. We test whether this is problematic using a large survey experiment on health and subjective well-being. We use Prelec's Bayesian truth serum to incentivize the experiment and defaults to introduce biases in responses. Without defaults, the data quality was good and incentives had no impact. With defaults, incentives reduced biases in the subjective well-being questions by inducing participants to spend more effort. Incentives had no impact on the health questions regardless of whether defaults were used.

Keywords: surveys; incentives; happiness; default bias.

JEL-codes: C83, C90.

[1] Erasmus School of Economics, Department of Applied Economics. Corresponding author: Aurelien Baillon, Erasmus University, PO box 1738, 3000 DR Rotterdam; Email: baillon@ese.eur.nl. We would like to thank Martijn Burger, Daniele Nosenzo, and Drazen Prelec for helpful comments. We also thank participants at the ESA World Meeting in Berlin, IMEBESS 2019, and the DFG PsychoEconomics workshop. This research was made possible by a Vidi grant (452-13-013) of the Netherlands Organization for Scientific Research (NWO). Internal Review Board approval: 2020/02/04-42584aba by ERIM – Section Experiments.

Hypothetical questions elicit hypothetical answers. Different versions of this fundamental critique of hypothetical choice scenarios have been voiced within and outside economics. Hypothetical choice is considered to have many problems (Wallis and Friedman 1942, Grether and Plott 1979, Bertrand and Mullainathan 2001). The first demonstrations of a mismatch between hypothetical and real decisions came from psychology (Edwards 1953, Slovic 1969). More recent evidence can also be found in economics (Holt and Laury 2002). However, opinions diverge how serious this mismatch is. The conclusions of overviews range from hypothetical and real choices being close (Beattie and Loomes 1997, Dohmen, et al. 2011, Falk, et al. 2019), to incentives moderately affecting behavior (Camerer and Hogarth 1999), to hypothetical choice being problematic (Hertwig and Ortmann 2001). Arguments against hypothetical choice (Guala 2005, p. 231ff., Read 2005, Bardsley, et al. 2010, p. 244ff) evoke lack of external validity and effort. Common opinion suggests using incentivized choices when possible (see, e.g., Camerer and Hogarth 1999, and , Bardsley, et al. 2010, for the case of experimental economics).

In fields like happiness or health economics, using real choices is typically impossible. Data involve self-reported feelings or health, which cannot be verified and, consequently, rewarded. Moreover, health economists are often interested in preferences over disease-related scenarios, like giving up life-years for improvements in health, and these usually cannot be elicited by real choices for obvious reasons. Participants are rewarded with a small fee, which is independent of their effort, and they are simply asked to report their "true answer" or "true preference". Whether they indeed answer truthfully cannot be verified and, consequently, the question whether these choices truly reflect their preferences or are subject to hypothetical bias cannot be answered.

Starting with Prelec's (2004) Bayesian truth-serum (BTS), several methods have been developed to incentivize unverifiable answers (Miller, Resnick and Zeckhauser 2005, Radanovic and Faltings 2013, Witkowski and Parkes 2012,

Baillon 2017, Cvitanić, et al. 2019). These methods assume that participants are Bayesians and interpret their own truthful answer as a private signal about others' answers. Consequently, participants expect that their own answer will be surprisingly common: more common than others predict. For example, people who are satisfied with their life will expect more people to be satisfied with their life than unsatisfied people do. This pattern has been confirmed in psychology (Ross, Greene and House 1977, Marks and Miller 1987).

The BTS asks respondents to do two things: to answer a question and to predict how common each possible answer will be. Respondents are rewarded for the accuracy of their prediction and they get a bonus if their own answer is surprisingly common. Prelec (2004) showed that truthful answering is the Bayesian Nash equilibrium that maximizes respondents' expected reward. Empirical evidence suggests that the BTS indeed increases truth-telling: John, Loewenstein and Prelec (2012) found that more psychologists admitted to questionable research practices when they were rewarded by the BTS and Weaver and Prelec (2013) showed that BTS-incentives reduce overclaiming of knowledge.

The central question of this paper is whether using hypothetical choices in surveys on subjective well-being and health is problematic and biases responses. To that end, we compare hypothetical answers and BTS-incentivized answers in tasks that are widely used to measure subjective well-being and health. We ran a large-scale, online experiment with 864 participants. Because the effect of incentives may depend on the quality of the data, we included treatments with and without defaults to exogenously influence data quality. Defaults strongly affect insurance and pension plan choice, and organ donation (Johnson, et al. 1993, Madrian and Shea 2001, Johnson and Goldstein 2003). They are used to 'nudge' people in desired

directions and they are the standard tool of behavioral insights teams.[2] In our study, defaults introduced a quantifiable bias, and we measured the extent to which incentives could reduce this bias.

We found that without defaults, incentives had no effect: data quality was good for both hypothetical and incentivized questions. This is consistent with Abeler, Nosenzo, and Raymond (2019), who show that people have a strong preference for truth-telling even when it is easy to lie. With defaults, however, incentives had an effect: they reduced the default bias, mainly because participants applied more cognitive effort. Incentives were only effective in the subjective well-being questions where responses were quick, and hence, more automatic. For the more complex health questions, which required more time to respond, incentives did not reduce the default bias. Our take-away message is that if researchers are careful about data quality and avoid defaults, and, probably, other biases, then hypothetical answers cause no problems in survey data. However, if data quality decreases, e.g. because of defaults, then incentives can help to reduce biases by stimulating effort.

1. Method

1.1. Method and participants

We ran a large-scale, online experiment involving 864 US residents. The sample involved no students, but it was not representative of the US population. Participants were recruited through the research platform Prolific, a UK-based alternative for Amazon's MTurk, specifically designed for research (Peer, et al. 2017, Palan and Schitter 2018). The survey was

[2] The Behavioral Insights team UK, for instance, recommends governments to use the power of defaults. See http://www.behaviouralinsights.co.uk/publications/east-four-simple-ways-to-apply-behavioural-insights/

administered using Qualtrics. **Table 1** summarizes the characteristics of the sample. The average earnings by participating in our experiment corresponded with an hourly wage rate of $12.76, more than twice the rate Prolific recommends as fair.

Variable	Control	Incentives	Defaults	Incentives & Defaults	Experiment
Number of observations	211	224	210	219	864
Mean age in years (std. dev.)	36.8 (11.9)	37.2 (10.7)	36.0 (10.6)	36.3 (10.5)	36.6 (10.9)
Gender:					
Female	49.3%	51.3%	51.4%	53.0%	51.3%
Male	48.8%	46.0%	47.6%	46.1%	47.1%
Other	1.4%	2.2%	1.0%	0.5%	1.3%
Not disclosed	0.5%	0.4%	0.0%	0.5%	0.3%
Income:					
Less than $25,000	21.3%	18.8%	14.8%	13.2%	17.0%
$25,000 to $49,999	20.4%	22.3%	24.3%	23.7%	22.7%
$50,000 to $64,999	9.5%	12.9%	12.9%	11.9%	11.8%
$65,000 to $104,999	17.5%	23.7%	21.9%	26.0%	22.3%
$105,000 or more	26.5%	20.1%	23.8%	21.9%	23.0%
Not disclosed	4.7%	2.2%	2.4%	3.2%	3.1%
Mean time taken in sec (std. dev.)	708 (301)	730 (303)	704 (338)	769 (368)	728 (329)
Mean earnings in $ (std. dev.)	2.58	2.58 (0.25)	2.58	2.58 (0.33)	2.58 (0.21)

Table 1: Descriptive statistics participants.

1.2. Survey

The survey consisted of three tasks: questions on subjective well-being, health, and recognition. Because measurements of subjective well-being and health typically employ hypothetical choices, we used these domains to explore the effects of incentives. We included the recognition task to replicate Weaver and Prelec (2013). Online Appendix B contains a complete transcript of the survey instructions.

Subjective well-being. We measured subjective well-being by asking participants for each of seven emotions (enjoyment, love, anger, pain, sadness, stress, worry) whether they had experienced it a lot the day before they took the survey. The exact wording was taken from the Gallup World Poll survey. Subjective well-being around the world is reported in the UN World Happiness Report (Helliwell, Layard and Sachs 2015).[3] The relationship between subjective well-being and economic variables, like income, has been documented in Kahneman et al. (2006), Kahneman and Deaton (2010), Deaton and Stone (2013), and Helliwell and Huang (2014).

Health. We asked participants two types of questions on health: the *standard gamble* (SG) and the *time trade-off* (TTO). The SG and the TTO are widely-used in health research to measure quality of life (Dolan, et al. 1996, Drummond, et al. 2015).

In the SG, participants chose between living in an impaired health state A for the rest of their life for sure and a risky treatment that gave a probability p of living in full health for the rest of their life and a probability $1 - p$ of immediate death. Hence, participants traded off improvements in quality of life against increases in the risk of immediate death. There were six SG questions in total, with the probability p varying from 0.15 to 0.90 in steps of 0.15. Each participant answered three randomly selected SG questions.

---

[3] The question wordings were taken from: https://media.gallup.com/dataviz/www/WP_Questions_WHITE.pdf (last access 04/05/2018). The emotions we used were a subset of those reported in the world data-base of happiness: https://worlddatabaseofhappiness.eur.nl/hap_cor/desc_hind.php?ind=1439 (last access 04/05/2018). We also elicited global life satisfaction via Cantril's Self-Anchoring Scale, which asks participants to evaluate their current life situation on a ladder ranging from 0, the worst possible life, to 10, the best possible life. Cantril's Self-Anchoring Scale was the only non-binary question in the survey and we, therefore, do not report it in this paper.

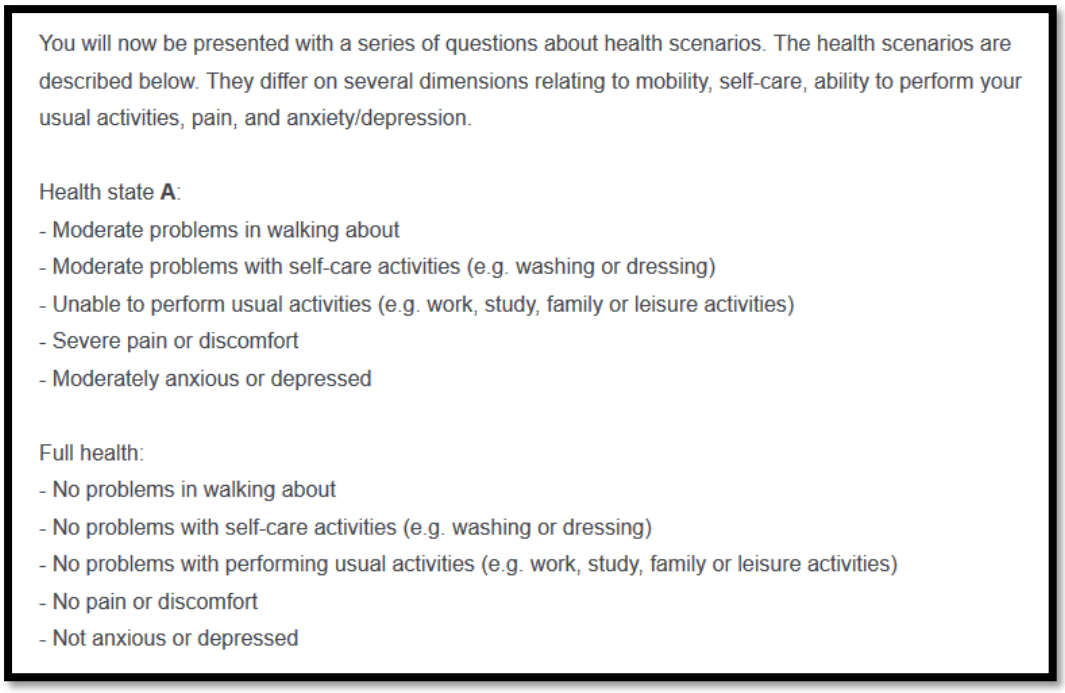

Figure 1: Screen-shot descriptions of health state A and full health used in experiment.

In the TTO questions, participants chose between 10 years in health state A (followed by death) and x years in full health. Hence, in these questions, participants traded off improvements in quality of life against reductions in life duration. There were six TTO questions with x varying between 1.5 and 9 years in steps of 1.5 years. Each participant answered three randomly chosen TTO questions.

Health state A and full health were described by the EuroQol EQ-5D-5L, which is common in health economics. The EQ-5D-5L describes health states in terms of five attributes: mobility, self-care, ability to perform usual activities, pain, and anxiety/depression. **Figure 1** shows the descriptions of health state A and full health.[4]

[4] https://euroqol.org/

Recognition. We used a shortened version of Weaver and Prelec's (2013) recognition questionnaire. We asked participants to indicate for each of twelve language concepts whether they recognized it. Eight concepts really existed, four were foils.[5] Our participants knew that some concepts were foils, but they did not know which ones or how many. People usually overclaim recognition. Weaver and Prelec (2013) showed that the BTS reduced overclaiming even when participants have incentives to overclaim.

For each question, participants were asked to answer it and to estimate the frequency with which a specific answer was selected by all participants. Frequency was measured by asking how many out of 100 randomly selected participants gave a specific answer. For example, in the recognition task participants were first asked whether they recognized the concept and then how many out of 100 randomly selected participants they thought had recognized this concept. We informed participants at the beginning of the experiment that they would be asked to estimate such frequencies. **Figure 2** and **Figure 3** give example screen-shots of the decision screens.

We used two orderings of the tasks. Participants either answered the questions in the order subjective well-being, health, recognition, or in the order recognition, health, subjective well-being. Each order had a 50% probability of being selected. Within tasks, the order of the questions was random.

[5] The existing concepts were alliteration, aphorism, eponym, euphemism, grapheme, hyperbole, interrogative, and limiting adjective. The foils were capacitance, interjunction, lexical shunt, and sentence stigma.

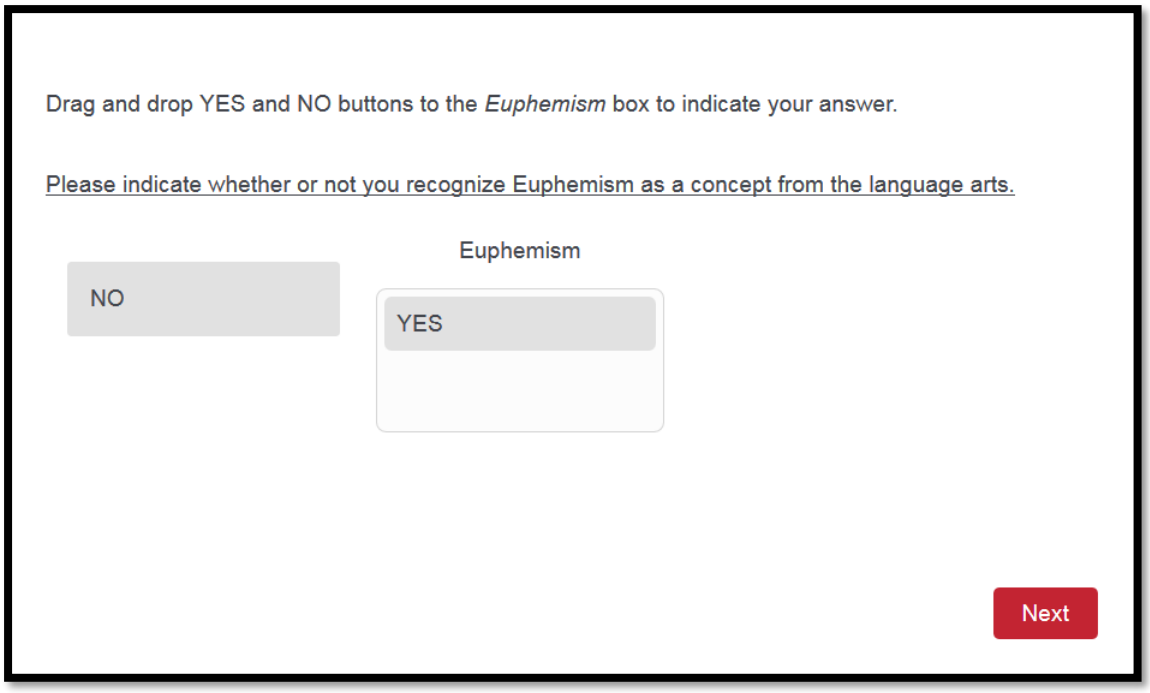

Figure 2: Screen-shot of a survey question. The example comes from the recognition questionnaire with a default.

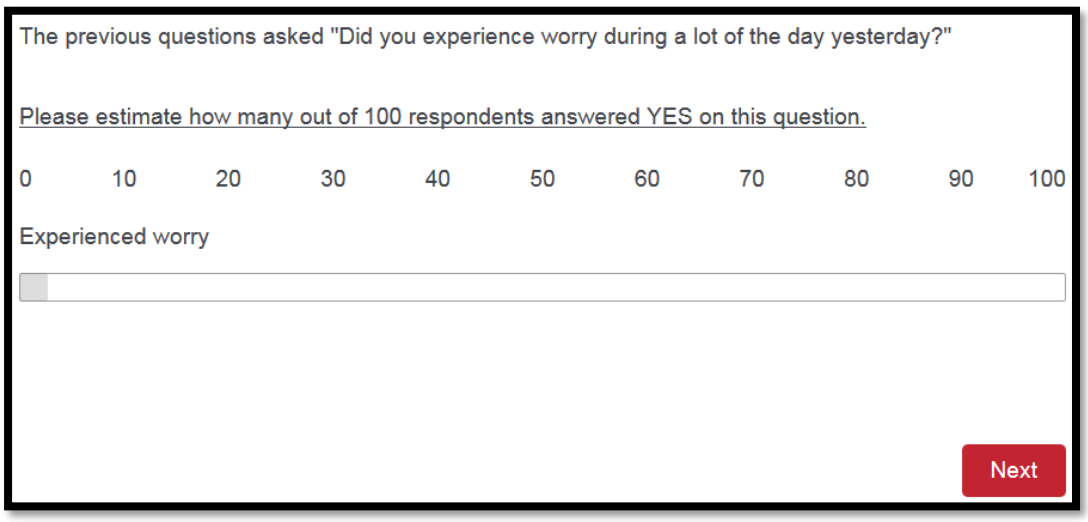

Figure 3: Screen-shot of the prediction task. The example comes from the subjective well-being questionnaire.

1.3. Experimental treatments

We randomly allocated the 864 participants to four different treatments, making sure that the number of participants per treatment was about equal.[6] The four treatments differed in terms of default answers and BTS incentives, as summarized in **Table 2**.

Treatment	Defaults set	BTS Incentives
Control		
Incentives		X
Defaults	X	
Defaults & Incentives	X	X

Table 2: Summary of experimental treatments.

Treatment *Control* (N=211) had no default answers and no BTS incentives. Participants received a flat participation fee of $2.58. They were asked to "*read all questions carefully and follow the on-screen instructions*", and to "*answer honestly and take care to avoid mistakes.*" The *Control* treatment reflects how surveys are usually administered.

Treatment *Defaults* (N=210) included defaults to bias participants' answers. Defaults worked through three channels (Johnson and Goldstein 2003). First, choosing the default option required less cognitive and physical effort than making an active choice (Samuelson and Zeckhauser 1988, Thaler and Sunstein 2008). We used an interface in which choices had to be made by dragging and dropping buttons into boxes. In treatments with defaults, the

[6] We aimed for 200 participants per treatment. Once these were reached, we stopped data collection, but allowed unfinished surveys to be completed. The target sample size was determined by a pilot in which 200 participants per treatment were enough to obtain significant results.

boxes contained the default answers. **Figure 2** gives an example**.** Changing the default answers required the effort of dragging answers from the box; sticking to the default required just one mouse-click. Second, the default might be perceived as a recommended action (McKenzie, Liersch and Finkelstein 2006). Third, defaults are the status quo and status quo bias may lead participants to choose it (Tversky and Kahneman 1991, Dinner, et al. 2011).

We chose the defaults to maximize their impact. Setting a "yes" default when most people choose "yes" anyhow, has little effect. We, therefore, used "no" as a default for positive emotions (for which most respondents typically answer "yes") and "yes" as a default for negative emotions (for which the common answer is "no"). In the health questions, the default was health state A, which was described as the status quo. In the recognition questionnaire, the default answer was "yes" to encourage overclaiming of recognition.

Treatments *Incentives* (N=224) and *Defaults & Incentives* (N=219) incentivized answers by the Bayesian Truth Serum (Prelec 2004). We followed the procedures of John, Loewenstein and Prelec (2012) and Weaver and Prelec (2013). Participants were informed that the BTS rewarded truth-telling, that it was invented by a professor from MIT, and that it had been published in SCIENCE, which is one of the most prestigious scientific journals. See Online Appendix B for the exact instructions. To keep stakes comparable across treatments, we restricted the BTS payments to the interval [$1.29, $3.87] with an average of $2.58. Participants were aware of this.

Table 1 shows the socio-demographic composition of each treatment group. We could not reject the null hypotheses that the treatment groups were similar in income, gender, and age, indicating that our randomization procedure had worked (all p-values > 0.17).[7]

[7] We ran a chi-square test for the categorical variables income (chi-square(22) = 27.456, p-value = 0.44) and gender (chi-square(3) = 0.486, p-value = 0.92). We ran pairwise

2. Results

2.1. Data quality

We first show that our data satisfy basic rationality requirements and are consistent with previous empirical evidence. We base this analysis on the two non-default treatments, *Control* and *Incentives*, to avoid the default bias and to be able to compare with previous studies.

Figure 4 shows the proportions of participants responding "Yes" to the various subjective well-being questions. We replicated two common observations (Kahneman and Deaton 2010, Helliwell, Layard and Sachs 2015): the *positive-negative experienced emotion gap*, more people report having experienced positive than negative emotions (in our study close to 70% reported positive emotions and less than 40% reported negative emotions) and the standard pattern that some emotions are reported more often than others. In particular, a clear majority of participants had experienced enjoyment (73%) and love (67%) the day before the survey. A smaller majority had experienced worry (55%) and stress (59%), common afflictions in modern societies. The more negative emotions sadness, anger, and pain were experienced less frequently (all < 34%).[8] Higher income was associated with more positive emotions, but not with negative emotions. This is in line with evidence that income increases life satisfaction, but not emotional well-being (Kahneman and Deaton 2010).[9]

comparison of means for the continuous variable age and time taken to complete the survey controlling for multiple testing via Tukey's honestly significant difference test. All results are reported in detail in Table A.8 and Table A.9 in Online Appendix A.

[8] For more details see **Figure A.1** in Online Appendix A.

[9] A more detailed account is in **Table A.1** in the Online Appendix A.

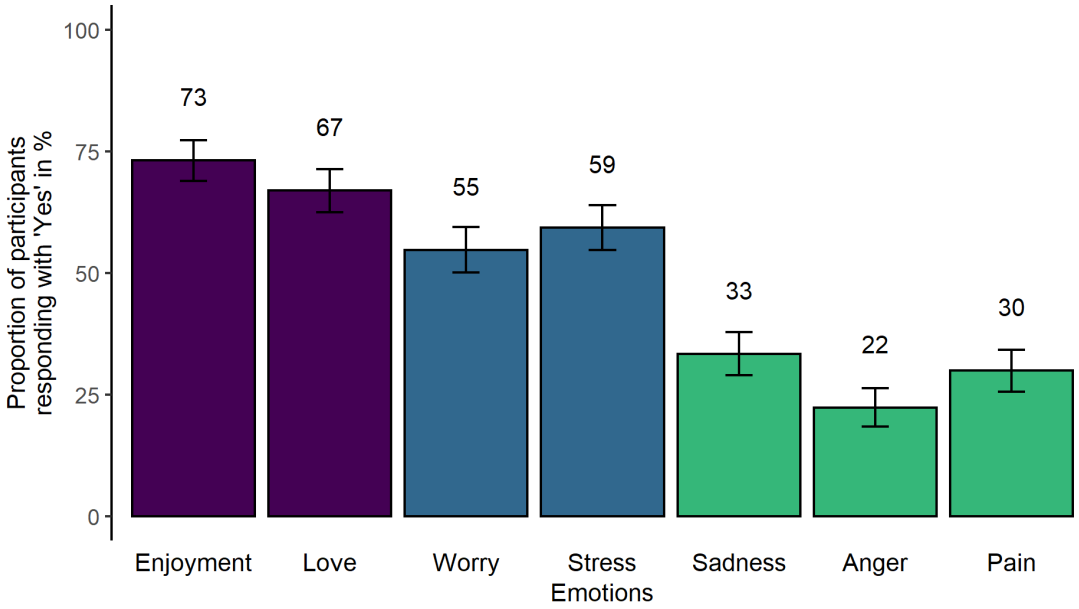

Figure 4: Proportion of participants responding "Yes" to the subjective well-being questions in Control and Incentives.

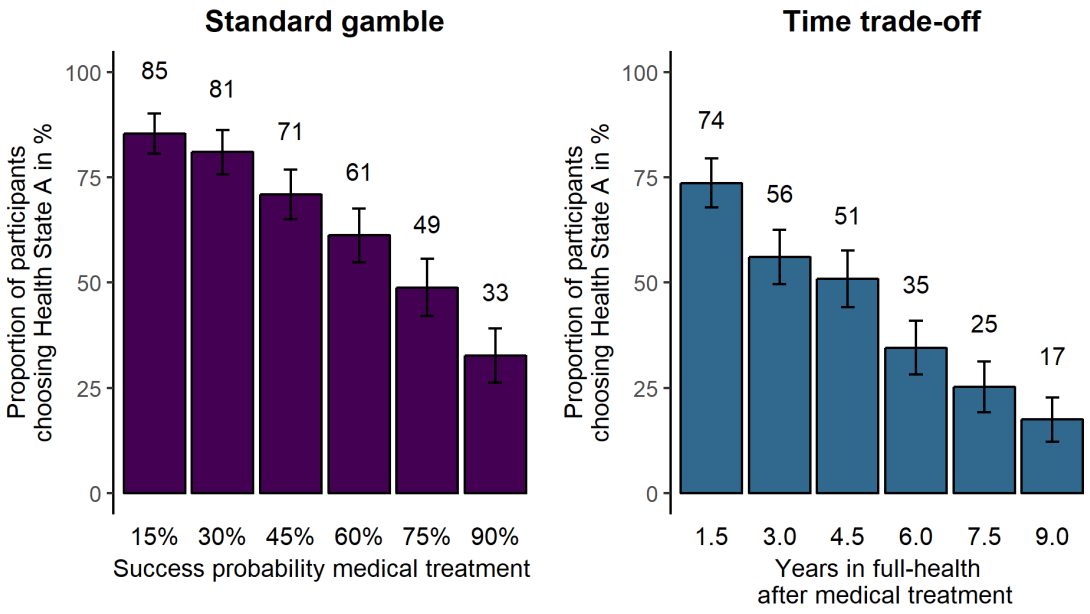

Figure 5: Proportions of participants in Control and Incentives choosing health state A in the SG as a function of the full-health probability (left panel) and in the TTO as a function of the number of years in full health (right panel).

Figure 5 shows that for health, the aggregate answers satisfied monotonicity: in the SG the risky treatment was chosen more frequently when the probability of full health increased and in the TTO participants chose the option full health more often when the number of years in full health increased. At the individual level, we observed that 402 out of 435 (93%) and 400 out of 435 (92%) participants satisfied monotonicity in the SG and in the TTO, respectively.[10]

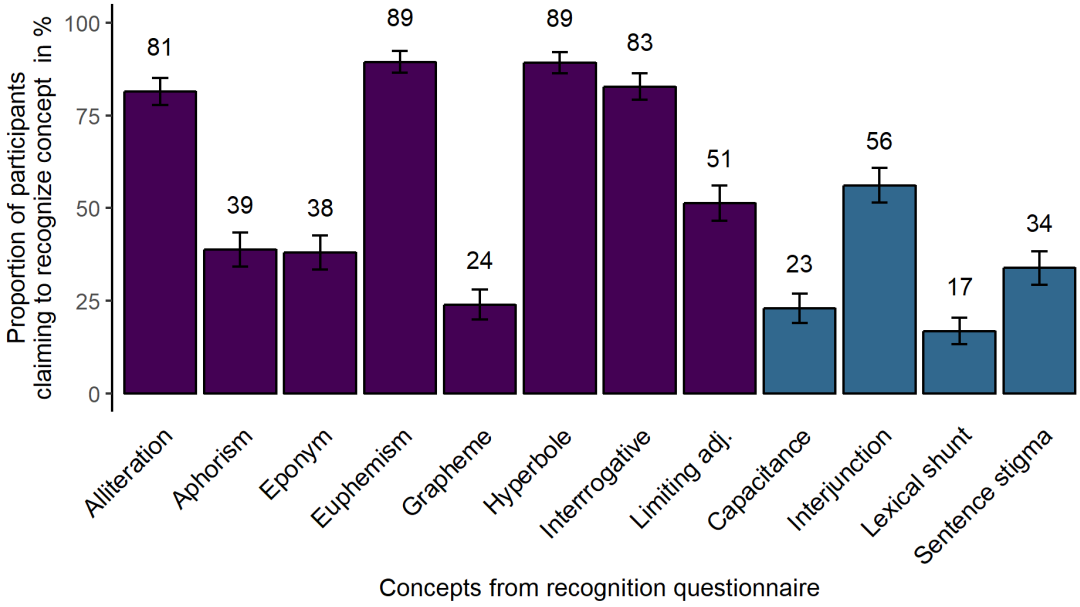

Figure 6: Proportion of participants claiming to recognize a concept in Control and Incentives. The last four concepts are foils.

Figure 6 shows that our participants recognized more extant concepts than foils (62% versus 32%) and that recognition rates for foils were well over 0%. These observations are compatible with previous studies (Paulhus, et al. 2003, Weaver and Prelec 2013). Well-known concepts like Alliteration, Euphemism, or Hyperbole were recognized frequently (> 80%). Less-known concepts like

[10] For more details, see Online Appendix A, model (3) in **Table A.2** and **Table A.3**.

Eponym and Grapheme were less often recognized (< 39%).[11] There is no clear correlation between socio-demographic variables and overclaiming for language concepts. Details are provided in the Online Appendix, **Table A.4**.

2.2. Effects of incentives

We study the general effect of incentives by comparing the *Incentives* and *Control* treatments. Overall, incentives had no effect. We found no effect of incentives across tasks either. We conjectured, for instance, that participants may be inclined to overreport socially desirable answers, "yes" for positive emotions and "no" for negative emotions, but probit regressions did not support this (Online Appendix A, **Table A.1**, models 3 and 4).

For SG and TTO, we conjectured that incentives might reduce the tendency to stick with the status quo (no medical treatment). However, we observed no such effects (Online Appendix A, **Table A.2** - model 2, and **Table A.3** - model 2). We also tested whether incentives made participants more sensitive to the attractiveness of the medical treatment (success probability of treatment in the SG or to the number of years in full health in the TTO), which would signal that they paid more attention to the questions. To do so, we included an interaction term between *Incentives* and the measures of attractiveness of treatment in the probit regressions. **Figure 7** displays the average marginal effects of *Incentives* on choosing no medical treatment as a function of the attractiveness of the medical treatment. If *Incentives* made respondents more sensitive to the attractiveness of the medical treatment, we would expect a downward trend. However, we did not observe this (see **Table A.5** in Online Appendix A for more details).

[11] More details are in **Figure A.4** in Online Appendix A.

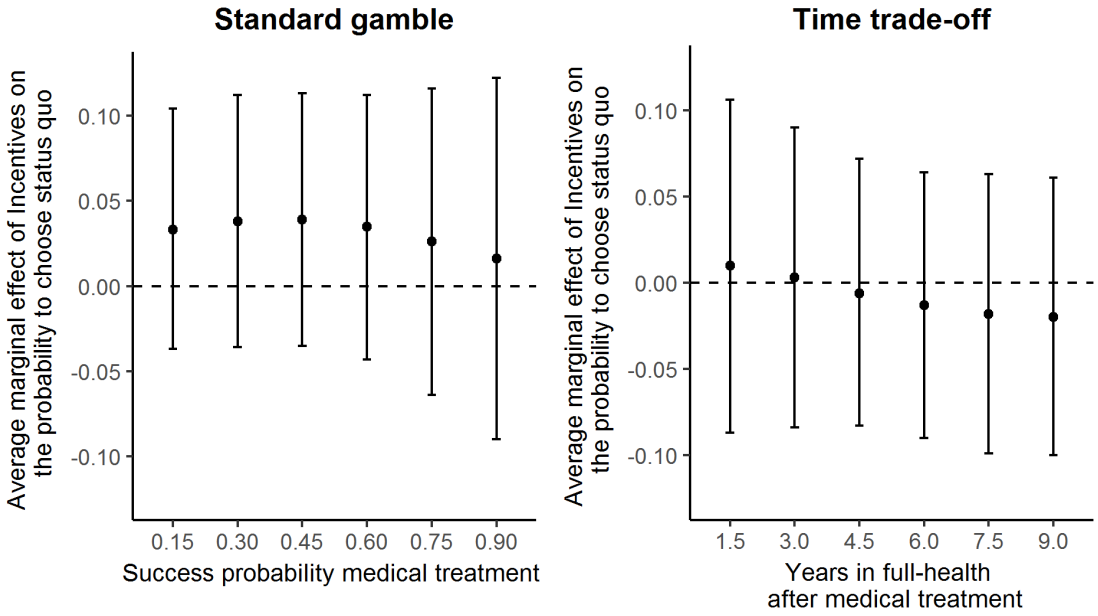

Figure 7: Average marginal effect of Incentives with respect to Control on the probability to the choose status quo (health state A) in the SG (left panel) and the TTO (right panel).

In the recognition questions, we hypothesized that incentives might reduce the recognition of foils. Again, we found no evidence for such an effect (Online Appendix A, **Table A.4** - model 4). This may be surprising, but it is consistent with the results of Weaver and Prelec (2013), who found that incentives only mattered when respondents were also paid per recognized item (prompting strong overclaiming), which we did not.

2.3. Effects of defaults

We refer to an option as the *default-favored option* if it was set as the default in the *Defaults* and *Defaults & Incentives* treatments. This terminology is slightly inaccurate as there were no defaults in the *Control* and *Incentives* treatments, but we believe it facilitates the presentation of the results. So, the default-favored options were admitting negative emotions and denying

positive emotions in the subjective well-being questions, the status quo (health state A) in the health questions, and recognition in the recognition task.

Figure 8 plots the proportion of choices for the default-favored option. The Figure shows that defaults worked: the default-favored options were chosen more often when defaults were included. Overall, the default bias was roughly 8 percentage points (ppt) without incentives (*Control* versus Defaults) and 5 ppt with incentives (*Incentives* versus *Defaults & Incentives*).

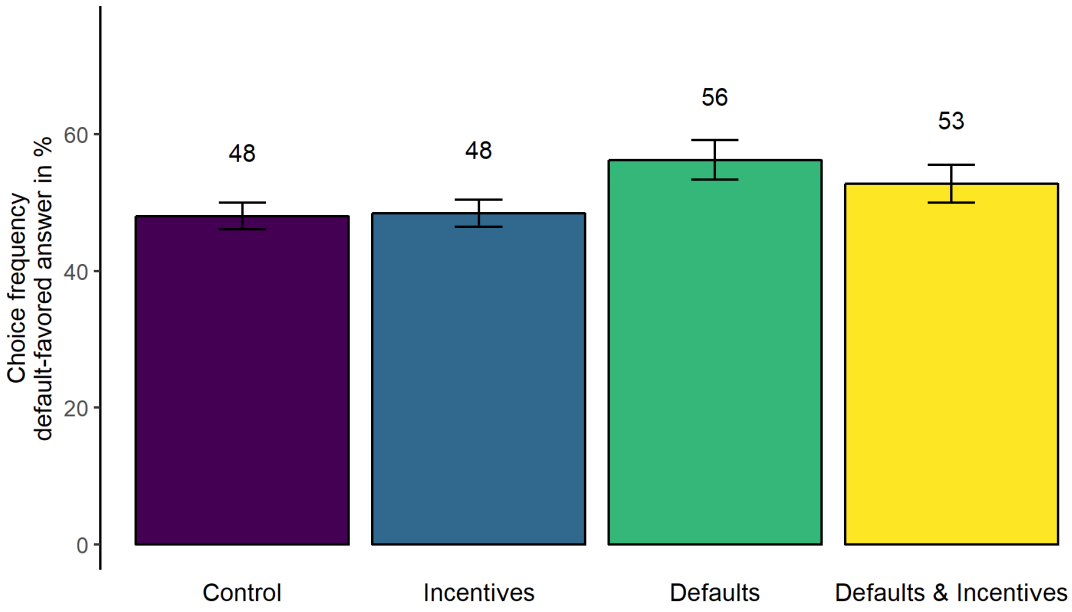

Figure 8: Frequency with which the default-favored option was chosen with 95% confidence intervals.

Figure 9 gives a more detailed account of the default bias by splitting up the choice proportions for the different tasks. The figure shows that the default bias was particularly strong for the subjective well-being questions (13 ppt without incentives and between 4 ppt and 8 ppt with incentives). In the recognition questions the default bias was weaker, ranging between 4 ppt and 8 ppt. In the SG, we observed no default bias at all. On the other hand, in the

TTO there was a clear default bias of 9 ppt with incentives and of 7 ppt without incentives.

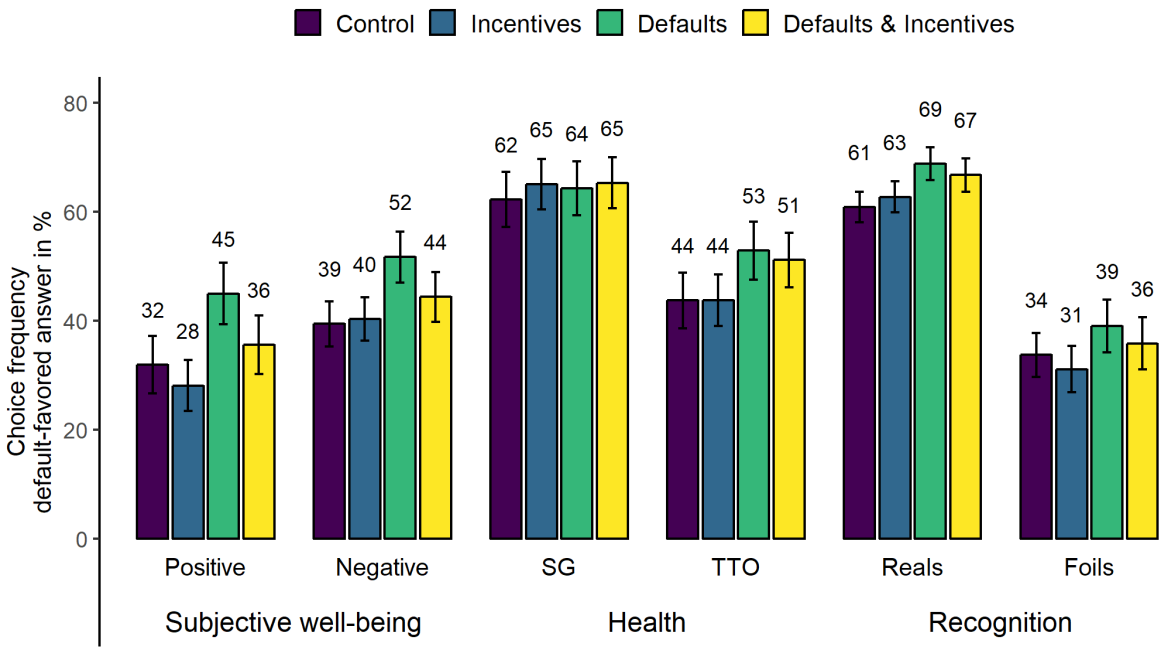

Figure 9: Frequencies with which the default-favored option was chosen in the various tasks with 95% confidence intervals.

All the above observations are confirmed by probit regressions with participant-clustered standard errors, and period (question order) and question fixed-effects. The results are in **Table 3**, and in **Table A.1** to **Table A.4** in Online Appendix A.

Model	(1)	(2)	(3)	(4)	(5)
Dep. Variable	Follow default-favored answer, in all models				
Question Domain	Pooled	SBJ. W-B	SG	TTO	REC
Incentives	0.016	0.005	0.038	0.001	0.021
	0.013	0.025	0.033	0.033	0.017
Defaults	0.091 ***	0.135 ***	0.025	0.103 **	0.080 ***
	0.016	0.028	0.034	0.035	0.018
Defaults & Incentives	0.058 ***	0.058 *	0.037	0.093 **	0.055 **
	0.016	0.027	0.034	0.034	0.018
Difference in estimated coefficients					
Defaults & Incentives vs Defaults	-0.033	-0.077 *	0.012	-0.010	-0.026
	0.018	0.030	0.033	0.035	0.020
Number of participants	864	864	864	864	864
Number of observations	21,600	6,048	2,592	2,592	10,368
Question FE	Yes	Yes	Yes	Yes	Yes
Period FE	Yes	Yes	Yes	Yes	Yes
Demographic controls	Yes	Yes	Yes	Yes	Yes
Participant clustered std. errors	Yes	Yes	Yes	Yes	Yes

*Table 3: Probit estimations with clustered standard errors at the participant level (reported below coefficient estimates in footnote size). Coefficient estimates represent average marginal effects. The dependent variable in all models takes the value 1 if a participant chooses the default-favored answer. All reported independent variables represent treatment dummies, Control being the baseline. The sample in model (1) includes all questions (Pooled). Models (2), (3), (4), and (5) restrict the sample to questions from the subjective well-being (SBJ W-B), the standard gamble (SG), the time trade-off (TTO), and the recognition questionnaire (REC), respectively. All models include fixed effects (FE) at the question and period level as well as demographic controls (age, and dummies for income and gender). We also report the estimated differences between the Defaults & Incentives and Defaults dummies. Significance coding: *** 0.1%, ** 1%, * 5%.*

2.4. Effect of incentives on default bias

Figure 8 shows that incentives reduced the default bias by 3 ppt: the default-favored option was less frequently chosen in the *Defaults & Incentives* treatment than in the *Defaults* treatment. The reduction of the default bias was

particularly strong in the subjective well-being questions, see **Figure 9**. In these questions, incentives reduced the default bias by 8 to 9 ppt. In the health and recognition questions, incentives had no effect on the default bias. Fixed-effect probit regressions confirm the above observations (see **Table 3**, and **Table A.1** to **Table A.4** in the Online Appendix A. The analysis of response times, which comes next, will shed light on the mechanism underlying the reduction of the default bias in the subjective well-being questions.

2.5. Response time

We recorded each participant's response time net of the time spent on the prediction task. Response time is often used as a proxy for effort, deliberation, and decision cost, which are typically unobservable (Moffat 2005, Bardsley, et al. 2010, Achtziger and Alós-Ferrer 2014, Rubinstein 2016). We conjectured that incentives would increase response time. Defaults could affect response time in opposite ways. They plausibly reduced the response time of those participants who wanted to choose the default option anyhow and of those participants who wished to go through the survey as fast as possible. Consequently, we expected a lower response time in treatments with defaults for those participants who preferred the default option. By contrast, defaults made reporting the non-default answer more time-consuming both physically (dragging out the default and dragging in the alternative) and cognitively (the dissonance between the participant's preference and the preselected option). To study these predictions, we regressed (the logarithm of) response time on the four treatments controlling for demographic variables.[12] **Figure 10** shows

[12] We used a panel generalized-least-squares approach with random effects at the individual level and cluster robust standard errors. This approach is standard in the literature on response time (Alós-Ferrer, et al. 2016, Moffatt 2016) . To account for outliers (a few raw response times were larger than 3000s), we winsorized response time at the 1% and 5% level

the results. It displays the deviation in response time from the Control treatment for the other three treatments. **Table A.7** in Online Appendix A gives additional results.

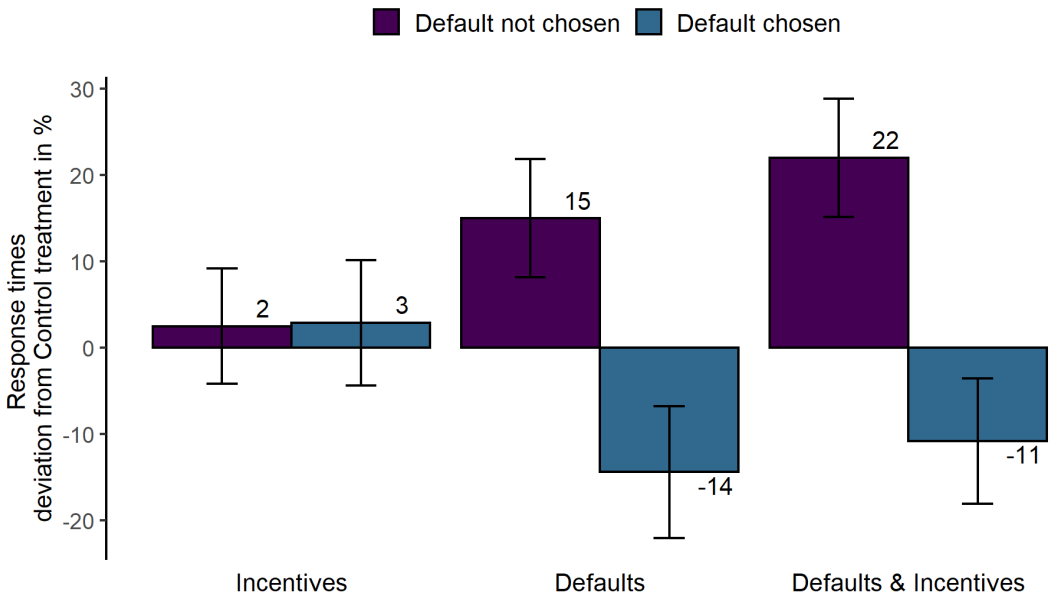

Figure 10: Percentage-deviation in response time with respect to the Control treatment for the different treatments.

Figure 10 shows that without defaults incentives did not affect response time. Defaults, on the other hand, had a clear impact on response time. In line with our predictions, participants spent on average 15% more time in *Defaults* than in *Control* when rejecting the default answer, but 14% less time when selecting the default answer (mean winsorized decision time was 10.87 seconds). When defaults were included, incentives increased the response time, but the effect was only significant when the default option was rejected. Our analysis suggests that participants were motivated, but that they could be

reducing maximum recorded response time to 74s and 33s, respectively. The results were very similar so we only report here the results with a 1% winsorization.

pushed to free ride on defaults. Incentives reduced this tendency to free ride. Participants spent on average between 3.6% and 7.0% more seconds answering the questions in *Defaults & Incentives* than in *Defaults*. These differences are net of any physical effort of using the drag-and-drop interface because defaults were set in both treatments. Hence, incentives may have convinced respondents to think more carefully and not to free ride on defaults.

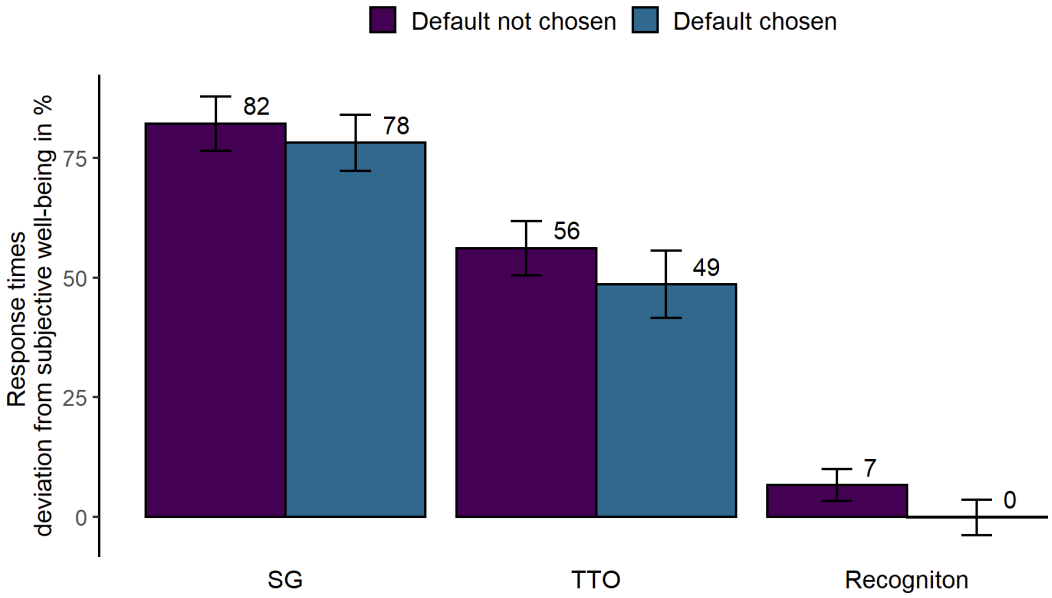

Figure 11: Percentage-deviation in response time with respect to the subjective well-being task for different tasks.

Response times did not differ between the *Incentives* and *Control* treatments Two opposing explanations might underlie this: participants are intrinsically motivated to answer surveys and do not need incentives or the incentives were too low. To further investigate these explanations, **Figure 11** compares the time spent on the different tasks. It shows the response time for the health and recognition tasks compared with the time spent on the subjective well-being questions. We expected that motivated participants would spend more time on the complex health questions, which involve trading off quality of life versus mortality risk and life duration, than on the other questions, which merely asked about having experienced an emotion or

recognizing a concept. This is indeed what we observed. We interpret this difference as evidence that participants took the questions seriously and that they were intrinsically motivated to answer the questions. We believe that this is the reason why we observed no effect of incentives without defaults, and not that incentives were too low. This observation is in line with our previous finding that answers were sensible (e.g., satisfying monotonicity in the health domain).

3. Discussion

Do incentives matter in survey data? Our results suggest that they do not if researchers are careful about data quality and avoid biases. If data quality was reduced, by defaults in our study, then incentives mattered: they reduced the default bias and increased the effort respondents spent on the tasks. This held in particular for relatively automatic tasks like reporting experienced emotions. For more complex tasks, like the health questions, we observed no effect of incentives regardless of whether defaults were used.

The absence of a significant effect of incentives when no defaults were used is consistent with previous evidence that even a flat compensation for 'work' can motivate participants to answer truthfully (see e.g., Gneezy and Rustichini 2000, Bardsley, et al. 2010, p. 250ff). An additional explanation might be that Prolific has a participant reputation system which is used for screening purposes. A lower reputation might mean fewer future invitations and this deterrence effect could have motivated the participants.

We could offer only small incentives (a few dollars) due to Prolific's regulations. The incentives were sufficiently large to affect behavior in the presence of defaults and our results suggest that they were adequate, but it would be interesting to explore the effect of larger incentives and whether they are worth the improvement in data quality. At least incentives never led to a decrease in data quality, not even for emotionally charged topics like health.

Previous research suggests that incentives can sometimes backfire if people feel they contribute to society (Titmuss 1970, Bleichrodt and Pinto Prades 2009). We found no evidence of such negative effects of incentives.

The absence of incentive effects in the more complex health questions was not caused by participants finding these questions so hard that they their preferences were incomplete. If so, we would expect a strong effect of defaults in these questions, as incomplete preferences are associated with status quo bias. The SG is commonly seen as the most difficult task in health state valuation and the response times in our study confirmed this. In the SG, we found no evidence of default bias.

In Prelec's (2004) BTS, participants have to perform an extra task, to predict the answers of other participants. Consequently, the BTS requires more cognitive effort. To test whether this extra cognitive effort affected data quality, we collected additional data for the *Control* and *Defaults* treatments without the prediction task.[13] Detailed results are in **Table A.6** in Online Appendix A. Excluding predictions had no effect in the *Control* treatment, but the effect of defaults was slightly weaker when participants were not asked to make predictions. This suggests that the extra prediction in the Control treatment was close to the maximum effort our participants were willing to make for the flat fee. Defaults were probably seen as an additional cognitive burden and then also adding the prediction task led to a minor tendency to go for the cognitively easier default-favored option. Recently, several papers have proposed Bayesian truth-telling mechanisms in which the prediction task is easier. For example, Baillon's (2017) Bayesian markets replace predictions by binary bets and in Cvitanic et al.'s (2019) choice-matching either only some

[13]The treatments with incentives needed predictions to compute the BTS scores. Alternatives are discussed later.

participants make predictions or predictions are simpler. It would be interesting to repeat our study with these mechanisms.

We did not explain to our participants why truth-telling is in their best interest in the BTS, as it requires knowledge of advanced mathematical concepts. Our implementation is sometimes called "intimidation" (Cvitanić, et al. 2019). We agree that this is a limitation of the BTS, particularly if the participants somehow mistrusted our claim that truth-telling was optimal. Weaver and Prelec (2013) implemented a version of the BTS in which they let participants experience the payment system and learn the strategy on their own. This approach is difficult to implement in surveys because 'live' computing of the BTS scores requires participants to answer simultaneously. Baillon's (2017) and Cvitanic et al.'s (2019) payment mechanisms involve simpler payment rules and computations, again making the case to repeat our study using these mechanisms.

4. Conclusion

Survey data are often used in research on subjective well-being and health. In these domains it is hard to use incentives, because responses are subjective and unverifiable. We used Prelec's (2004) BTS to explore whether this lack of incentives is problematic and hypothetical bias occurs. We found no evidence for hypothetical bias when good quality data are collected and biases are avoided. Avoiding biases seems more important than introducing incentives in surveys. Our participants were motivated, their responses satisfied all basic rationality requirements, and (without defaults) they were not affected by incentives. However, their answers were affected by defaults. Incentives reduced the default bias and increased effort in the subjective well-being questions. In the more complex health questions incentives had no effect regardless of whether defaults were used or not.

5. References

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond. 2019. "Preferences for Truth-Telling." *Econometrica* 1115-1153.

Achtziger, Anja, and Carlos Alós-Ferrer. 2014. "Fast or Rational? A Response-Times Study of Bayesian Updating." *Management Science* 60 (4): 923-938.

Alós-Ferrer, Carlos, Đura-Georg Granić, Johannes Kern, and Alexander K. Wagner. 2016. "Preference Reversals: Time and Again." *Journal of Risk and Uncertainty* 52 (1): 65–97.

Baillon, Aurelien. 2017. "Bayesian Markets to Elicit Private Information." *Proceedings of the National Academy of Sciences* 114.30: 7958-7962.

Bardsley, Nicholas, Robin Cubitt, Graham Loomes, Peter Moffatt, Chris Starmer, and Robert Sugden. 2010. *Experimental Economics: Rethinking the Rules.* Princeton and Oxford: Princeton University Press.

Beattie, Jane, and Graham Loomes. 1997. "The Impact of Incentives Upon Risky Choice." *Journal of Risk and Uncertainty* 14: 155-168.

Bertrand, Marianne, and Sendhil Mullainathan. 2001. "Do People Mean What They Say? Implications for Subjective Survey Data." *The American Economic Review* 91 (2): 67-72.

Bleichrodt, Han, and Jose L. Pinto Prades. 2009. "New Evidence of Preference Reversals in Health Utility Measurement." *Health Economics* 18 (6): 713-726.

Bleichrodt, Han, Jose Luis Pinto Prades, and Jose Maria Abellan-Perpiñan. 2003. "A Consistency Test of the Time Trade-off." *Journal of Health Economics* 1037-1052.

Camerer, Colin F., and Robin M. Hogarth. 1999. "The Effects of Financial Incentives in Experiments: A Review and a Capital-Labor-Production Framework." *Journal of Risk and Uncertainty* 19: 7-42.

Cvitanić, Jakša, Drazen Prelec, Blake Riley, and Benjamin Tereick. 2019. "Honesty via Choice Matching." *American Economic Review: Insights* 179-192.

Dawes, Robyn M. 1989. "Statistical Criteria for Establishing a Truly False Consensus Effect." *Journal of Experimental Social Psychology* 25 (1): 1-17.

Deaton, Angus, and Arthur A. Stone. 2013. "Two Happiness Puzzles." *American Economic Review* 3: 591-597.

Dinner, Isaac, Eric J. Johnson, Daniel G. Goldstein, and Kaiya Liu. 2011. "Partitioning Default Effects: Why People Choose Not to Choose." *Journal of Experimental Psychology: Applied* 17 (4): 332-341.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. 2011. "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences." *Journal of the European Economic Association* 9 (3): 522-550.

Dolan, P., C. Gudex, P. Kind, and A. Williams. 1996. "Valuing Health States: A Comparison of Methods." *Journal of Health Economics* 15 (2): 209-231.

Drummond, Michael F., Mark J. Sculpher, Karl Claxton, Greg L. Stoddart, and George W. Torrance. 2015. *Methods for the Economic Evaluation of Health Care Programmes.* Oxford, New York, Toronto: Oxford University Press.

Edwards, Ward. 1953. "Probability Preferences in Gambling." *The American Journal of Psychology* 66: 349-364.

Falk, Armin, Becker, Aanke, Thomas Dohmen, Benjamin Enke, Huffman, David, and Uwe Sunde. 2019. "Global Evidence on Economic Preferences." *The Quarterly Journal of Economics* 133 (4): 1645-1692.

Gneezy, Uri, and Aldo Rustichini. 2000. "Pay Enough or Don't Pay at All." *The Quartely Jounal of Economics* 115: 791-810.

Grether, David M., and Charles R. Plott. 1979. "Economic Theory of Choice and the Preference Reversal Phenomenon." *American Economic Review* 69: 195-207.

Guala, Francesco. 2005. *The Methodology of Experimental Economics.* New York: Cambridge University Press.

Helliwell, John F., and Haifang Huang. 2014. "New Measures of the Costs of Unemployment: Evidence From the Subjective Well-Being of 3.3 Million Americans." *Economic Inquiry* 52 (4): 1485-1502.

Helliwell, John, Richard Layard, and Jeffrey Sachs. 2015. *World Happiness Report 2015.* New York:: Sustainable Development Solutions Network.

Hertwig, Ralph, and Andreas Ortmann. 2001. "Experimental Practices in Economics: A Methodological Challenge for Psychologist?" *Behavioral and Brain Sciences* 24: 433-451.

Holt, Charles A., and Susan Laury. 2002. "Risk Aversion and Incentive Effects." *The Amreican Economic Review* 92 (5): 1644-1655.

John, Leslie K., George Loewenstein, and Drazen Prelec. 2012. "Measuring the Prevalence of Questionable Research Practices with Incentives for Truth-telling." *Psychological Science* 23 (5): 524-532.

Johnson, Eric J, John Hershey, Jacqueline Meszaros, and Howard Kunreuther. 1993. "Framing, Probability Distortions, and Insurance Decisions." *Journal of Risk and Uncertainty* 7: 35-51.

Johnson, Eric J., and Daniel Goldstein. 2003. "Do Defaults Save Lives?" *Science* 302: 1338-1339.

Kahneman, Daniel, Alan B. Krueger, David Schkade, Norbert Schwarz, and Arthur A. Stone. 2006. "Would You Be Happier If You Were Richer? A Focusing Illusion." *Science* 312: 1908-1910.

Kahneman, Daniel, and Angus Deaton. 2010. "High Income Improves Evaluation of Life But Not Emotional Well-Being." *Proceedings of the National Academy of Sciences* 107 (38): 16489-16493.

Madrian, Brigitte C., and Dennis F. Shea. 2001. "The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior." *The Quarterly Journal of Economics* 116 (4): 1149–1187.

Marks, Gary, and Norman Miller. 1987. "Ten Years of Research on the False-Consensus Effect." *Psychological Bulletin* 102 (1): 72.

McKenzie, Craig R. M., Michael J. Liersch, and Stacey R. Finkelstein. 2006. "Recommendations Implicit in Policy Defaults." *Psychological Science* 17 (5): 414-420.

Miller, Nolan, Paul Resnick, and Richard Zeckhauser. 2005. "Eliciting Informative Feedback: The Peer Prediction Method." *Management Science* 51 (9): 1359-1373.

Moffat, Peter G. 2005. "Stochastic Choice and the Allocation of Effort." *Experimental Economics* 8 (4): 369-388.

Moffatt, Peter G. 2016. *Experimetrics: Econometrics for Experimental Economics.* Macmillan International Higher Education.

Palan, Stefan, and Christian Schitter. 2018. "Prolific.ac - A Subject Pool for Online Experiments." *Journal of Behavioral and Experimental Finance* forthcoming.

Paulhus, Delroy L., P. D. Harms, M. Nadine Bruce, and Daria C. Lysy. 2003. "The Over-Claiming Technique: Measuring Self-Enhancement Independent of Ability." *Journal of Personality and Social Psychology* 84 (4): 890-904.

Peer, Eyal, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. "Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research." *Journal of Experimental Social Psychology* 70: 153-163.

Prelec, Drazen. 2004. "A Bayesian truth serum for subjective data." *Science* 306: 462-466.

Radanovic, Goran, and Boi Faltings. 2013. "A Robust Bayesian Truth Serum for Non-Binary Signals." *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI' 13).* 833-839.

Read, Daniel. 2005. "Monetary Incentives, What Are They Good For?" *Journal of Economic Methodology* 12 (2): 265-276.

Ross, Lee, David Greene, and Pamela House. 1977. "The "False Consensus Effect". An Egocentric Bias in Social Perception and Attribution Processes." *Journal of Experimental Social Psychology* 13 (3): 279-301.

Rubinstein, Ariel. 2016. "A Typology of Players: Between Instinctive and Contemplative." *Quarterly Journal of Economics* 131: 859-890.

Samuelson, William, and Richard Zeckhauser. 1988. "Status Quo Bias in Decision Making." *Journal of Risk and Uncertainty* 1: 7-59.

Slovic, Paul. 1969. "Differential Effects on Real versus Hypothetical Payoff on Choices Among Gambles." *Journal of Experimental Psychology* 80: 434-437.

Thaler, Richard H., and Cas R. Sunstein. 2008. *Nudge: Improving Decisions About Health, Wealth, and Happiness.* New Haven: Yale University Press.

Titmuss, Richard. 1970. *The Gift Relationship.* London: Allen and Unwin.

Tversky, Amos, and Daniel Kahneman. 1991. "Loss Aversion in Riskless Choice: A Reference-Dependent Model." *The Quarterly Journal of Economics* 106 (4): 1039-1061.

Wallis, W. Allen, and Milton Friedman. 1942. "The Empirical Derivation of Indifference Functions." In *Studies in Mathematical Economics and Econometrics in Memory of Henry Schultz,* edited by Lange O., McIntyre F. and Yntema T. O., 175-189. Chicago: University of Chicago Press.

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

Witkowski, Jens, and David C. Parkes. 2012. "A Robust Bayesian Truth Serum for Small Populations." *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI' 12).* 1492-1498.

Online Appendix

6. Appendix A: Additional Analyses

This appendix extends the data analysis as reported in the main text. Whereas we decided to sacrifice comprehensiveness for clarity and conciseness in reporting our main results, we aim to demonstrate here that our main results are robust and obtain under a broader set of analyses. We will follow the same empirical strategy that we have adopted in **Section 2**. Specifically, we use probit models to estimate the average marginal effect of various variables of interest on the probability to choose default-favored answers. Independent observations are taken at the participant-question level. Therefore, all estimation results are based on participant clustered standard errors. We also include various fixed effects to account for variation at the question-level and the question round-level. As demographic controls, we include the age measured in years, and dummies representing each possible answer category for the income question and gender question (including non-disclosure).

Table A.1 reports the corresponding regression results for the subjective well-being part. In model (1), the dummy *Defaults set* takes the value 1 if a default was set for the question. Pooled over treatments, defaults increased the probability to choose default-favored answers by 8.6 percentage points (ppt). Defaults were thus highly effective in this question domain. Treatment differences in the power of defaults are analyzed in models (2) to (4). Model (2) is the same as Model (3) in **Table 3** and already discussed in the main text. Models (3) and (4) break down the data into negative and positive emotions, respectively. As evidenced by the difference in estimated coefficients between *Defaults & Incentives* and *Defaults*, BTS-incentives reduced the default bias by 7.1ppt for negative emotions and by 9.4ppt for positive emotions. Finally, Models (5) and (6) report the results of our association-analysis between participants' characteristics and the probability to admit negative and positive feelings. Note the change in the dependent variable, which takes the value 1 if a participant admits an emotion (answers with YES). We restricted the sample to non-defaults treatments, we excluded participants with undisclosed personal characteristics, and we introduced slight changes in the way we treated the demographic variables. These changes were necessary to be able to compare our results with the existing literature. In particular, we transformed

gender into the *Male* dummy with value 1 if a participant is male and treated the 9 different income categories as a continuous variable in ascending order (*Income*). A one-unit change in the *Income* variable thus represents the impact of going up one income category on the probability to admit having experienced an emotion. *Age* still measures participants' age in years. As reported in the main text, we find a positive association between *Income* and having experienced positive emotions, but not for negative emotions. Moving up one income category on average increased the probability of reporting positive emotions by 3.4ppt. In comparison, incentives reduced the default bias on positive emotions by 9.4ppt, three times as much. This illustrates the economic significance of the reduction of default bias by using incentives. A more detailed account of the underlying data is in **Figure A.1,** which plots the emotion-admittance rates for each of the seven emotions split out by treatment.

Table A.2 and **Table A**.3 present the regression results for the health domain. We find large differences in the effectiveness of defaults between SG and TTO. Model (1) in both tables includes the dummy *Default set* which takes the value 1 if a default was set in a question. In the SG, defaults had no significant effect. In the TTO, we found a significant and positive effect of defaults: the probability to choose the default option health state A increased by 9.8ppt. Model (2) in **Table A.2** and **Table A**.3 corresponds to model (4) and model (5) in **Table 3** and we refer the reader to the main text for their discussion. Model (3) in the tables presents the association analysis as reported in the main text in **Section 2.1.** The exogenous question parameters significantly impacted behavior in the expected direction. In the SG, increasing the success probability of the medical treatment by one ppt decreased the choice probability for health state A by 0.659 ppt. In the TTO, increasing the number of years spent in full health after the medical treatment by one year decreased the probability of choosing health state A by 7.1 ppt.

In the TTO, older participants and males were more likely to choose health state A. An increase in age by one year increased the probability to choose health state A by 0.4 ppts. Similarly, male participants had a 9.7ppt higher probability to choose health state A than female participants. The demographic variables did not significantly affect the SG. **Figure A.2** and **Figure** A.3 provide a more detailed account of the underlying data and break down the relative choice frequency for health state A to the question and treatment level. As can be seen, with a minor exception, the choice frequencies in favor of health state A are monotonically decreasing in the attractiveness of the medical treatment across all experimental treatments.

Table A.4 presents the results for the recognition questionnaire. Again, defaults were highly effective: they increased the probability to recognize concepts by 5.6 ppt (*Defaults set* in Model (1)). Models (3) and (4) restrict the sample to existing concepts (Reals) and non-existing concepts (Foils), respectively. There were no substantial differences between reals and foils. See also **Figure A.4**. As we have explained in the main text, the literature has found no consistent socio-demographic correlates of overclaiming propensities. Models (5) and (6), nevertheless, show that in our study males were 7.2ppt and 12ppt more likely to recognize Reals and Foils, respectively. Income was also positively associated with recognition rates. As income tends to be positiviely correlated with eduction, this finding may reflect that higher educated individuals find it harder to admit not recognizing a concept.

The main text reported that we found no differences in choice probabilities between *Control* and I*ncentives* across all success probabilities in the SG and all years in full health in the TTO. **Table A.5** presents the underlying probit estimates that we used for **Figure 7** in the main text. Models (1) and (2) are analogous to model (3) in **Table A.2** and **Table** A.3, except that we included interaction terms between the *Incentives* dummy and variables that measure the success probabilities and years in full health.

Finally, we ran two additional treatments, replicating *Control* and *Defaults* without the prediction task, to investigate whether the prediction task influenced our results (as it is integral to the BTS, we could not run the *Incentives* and *Incentives&Defaults* treatments without the prediction task). **Table A.6** presents the corresponding relative choice frequencies for default-favored answers. We found systematic differences between the *Control* and *Control without prediction* treatments. With defaults, the prediction task increased the default bias in three out of the four question domains. Two-sided Mann-Whitney-U tests and two-sided t-tests, both with Holm-Bonferroni correction, corroborate these observations statistically. As discussed in the main text, our findings suggest that participants in *Control* were close to the maximum effort they were willing to exert for the flat fee and that the prediction task itself imposed a small cost, which affected the more effortful *Default* treatment.

Model	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Variable	Follow default favored answer (1) - (4)				Admitt emotion (5) - (6)	
Emotions	All	All	Negative	Positive	Negative	Positive
Defaults set	0.093 ***	-	-	-	-	-
	0.019					
Incentives	-	0.005	0.021	-0.033	-	-
		0.025	0.029	0.034		
Defaults	-	0.135 ***	0.133 ***	0.145 ***	-	-
		0.028	0.031	0.034		
Defaults & Incentives	-	0.058 *	0.062 *	0.052	-	-
		0.027	0.031	0.037		
Age	-	-	-	-	0.000	0.000
					0.001	0.001
Male	-	-	-	-	0.025	0.016
					0.030	0.036
Income	-	-	-	-	0.004	0.034 ***
					0.006	0.007
Difference in estimated coefficients						
Defaults & Incentives vs Defaults	-	-0.077 *	-0.071 *	-0.094 *	-	-
		0.030	0.032	0.039		
Number of participants	864	864	864	864	420	420
Number of observations	6,048	6,048	4,320	1,728	2,100	840
Question FE	Yes	Yes	Yes	Yes	Yes	Yes
Period FE	Yes	Yes	Yes	Yes	Yes	Yes
Demographic controls	Yes	Yes	Yes	Yes	Reported	Reported
Participant clustered std. errors	Yes	Yes	Yes	Yes	Yes	Yes

*Table A.1: Probit estimations for the subjective well-being part with clustered standard errors at the participant level (reported below coefficient estimates in footnote size). Coefficient estimates represent average marginal effects. The dependent variable in models (1) to (4) takes the value 1 if a participant chooses the default-favored answer. The dependent variable in models (5) and (6) takes the value 1 if a participant admits an emotion. For positive emotions the default was 'NO', for negative emotions the default was 'YES'. The independent variable 'Defaults set' is a dummy and takes the value 1 if defaults are set. We include further dummies that represent our treatments, taking 'Control' as the baseline. 'Age' measures participants' age in years, 'Male' is a dummy with value 1 if participant is male, and 'Income' represents the 9 different income categories for household income (ascending order, treated as continuous variable). The samples in model (1) and (2) include all questions. Models (3) and (5) restrict the sample to negative emotions. Models (4) and (6) restrict the sample to positive emotions. Models (5) and (6) restrict the sample to non-default treatments excluding participants with undisclosed personal characteristics. All models include fixed effects (FE) at the question and period level. We also report the estimated differences in the effect of defaults due to incentives / cognitive load. Significance coding: *** 0.1%, ** 1%, * 5%.*

Model	(1)	(2)	(3)
Dep. Variable	Follow default-favored answer, all models		
Defaults set	0.012	-	-
	0.024		
Incentives	-	0.038	-
		0.033	
Defaults	-	0.025	-
		0.034	
Defaults & Incentives	-	0.037	-
		0.034	
Success probability medical treatment	-	-	-0.659 ***
			0.042
Age	-	-	0.002
			0.002
Male	-	-	0.009
			0.025
Income	-	-	0.010
			0.007
Difference in estimated coefficients			
Defaults & Incentives vs Defaults	-	0.012	-
		0.033	
Number of participants	864	864	420
Number of observations	2,592	2,592	1,260
Question FE	Yes	Yes	No
Period FE	Yes	Yes	Yes
Demographic controls	Yes	Yes	Reported
Participant clustered std. errors	Yes	Yes	Yes

*Table A.2: Probit estimations for the standard gamble with clustered standard errors at participant level (reported below coefficient estimates in footnote size). Coefficient estimates represent average marginal effects. The dependent variable in all models takes the value 1 if a participant chooses the default-favored answer. The default was always set to favor health state A. The independent variable 'Defaults set' is a dummy and takes the value 1 if defaults are set. We also include further dummies representing treatments, with 'Control' as the baseline. 'Success probability medical treatment' represents the success probability of the medical treatment from 0.00 to 1.00, 'Age' measures participants' age in years, 'Male' is a dummy with value 1 if participant is male, and 'Income' represents the 9 different income categories for household income (ascending order, treated as continuous variable). The samples in model (1) and (2) include all questions. Models (3) restricts the sample to non-default treatments excluding participants with undisclosed personal characteristics. All models include fixed effects (FE) at period level. We also report the estimated differences in the effect of defaults due to incentives / cognitive load. Significance coding: *** 0.1%, ** 1%, * 5%.*

Model	(1)	(2)	(3)
Dep. Variable	Follow default-favored answer, all models		
Defaults set	0.098 ***	-	-
	0.024		
Incentives	-	0.001	-
		0.033	
Defaults	-	0.103 **	-
		0.035	
Defaults & Incentives	-	0.093 **	-
		0.034	
Years in full health medical treatment	-	-	-0.071 ***
			0.004
Age	-	-	0.004 *
			0.002
Male	-	-	0.097 **
			0.034
Income	-	-	0.002
			0.007
Difference in estimated coefficients			
Defaults & Incentives vs Defaults	-	-0.010	-
		0.035	
Number of participants	864	864	420
Number of observations	2,592	2,592	1,260
Question FE	Yes	Yes	No
Period FE	Yes	Yes	Yes
Demographic controls	Yes	Yes	Reported
Participant clustered std. errors	Yes	Yes	Yes

*Table A.3: Probit estimations for the time trade-off part clustered standard errors at participant level (reported below coefficient estimates in footnote size). Coefficient estimates represent average marginal effects. The dependent variable in all models takes the value 1 if a participant chooses the default-favored answer. The default was always set to favor health state A, i.e. no medical treatment. The independent variable 'Defaults set' is a dummy and takes the value 1 if defaults are set. We also include further dummies representing treatments, with 'Control' as the baseline. 'Years in full health medical treatment' represents the life-time in full health after the medical treatment, from 1.5 to 9.0 years. Age' measures participants' age in years, 'Male' is a dummy with value 1 if participant is male, and 'Income' represents the 9 different income categories for household income (ascending order, treated as continuous variable). The samples in model (1) and (2) include all questions. Models (3) restricts the sample to non-default treatments excluding participants with undisclosed personal characteristics. All models include fixed effects (FE) at period level. We also report the estimated differences in the effect of defaults due to incentives / cognitive load. Significance coding: *** 0.1%, ** 1%, * 5%.*

Model	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Variable	Follow default-favored answer, all models					
Concepts	Pooled	Pooled	Reals	Foils	Reals	Foils
Defaults set	0.056 ***	-	-	-	-	-
	0.013					
Incentives	-	0.021	0.032	-0.006	-	-
		0.017	0.020	0.026		
Defaults	-	0.080 ***	0.090 ***	0.059 *	-	-
		0.018	0.020	0.027		
Defaults & Incentives	-	0.055 **	0.064 **	0.031	-	-
		0.018	0.020	0.026		
Age	-	-	-	-	0.000	0.001
					0.001	0.001
Male	-	-	-	-	0.072 ***	0.157 ***
					0.020	0.021
Income	-	-	-	-	0.014 ***	0.036 ***
					0.004	0.005
Difference in estimated coefficients						
Defaults & Incentives vs Defaults	-	-0.026	-0.025	-0.029	-	-
		0.020	0.020	0.030		
Number of participants	864	864	864	864	420	420
Number of observations	10,368	10,368	6,912	3,456	3,360	1,680
Question FE	Yes	Yes	Yes	Yes	Yes	Yes
Period FE	Yes	Yes	Yes	Yes	Yes	Yes
Demographic controls	Yes	Yes	Yes	Yes	Reported	Reported
Participant clustered std. errors	Yes	Yes	Yes	Yes	Yes	Yes

*Table A.4: Probit estimations for the recognition questionnaire with clustered standard errors at participant level (reported below coefficient estimates in footnote size). Coefficient estimates represent average marginal effects. The dependent variable in all models takes the value 1 if a participant chooses the default-favored answer. The default was always set to YES recognizing each concept. The independent variable 'Defaults set' is a dummy and takes the value 1 if defaults are set. We also include further dummies representing treatments, with 'Control' as the baseline. 'Age' measures participants' age in years, 'Male' is a dummy with value 1 if participant is male, and 'Income' represents the 9 different income categories for household income (ascending order, treated as continuous variable). The samples in model (1) and (2) include all concepts (Pooled). Models (3) and (5) restrict the sample to existing concepts (Reals). Model (4) and (6) restrict the sample to non-existing concepts (Foils). Model (5) and (6) restrict the sample to non-default treatments excluding participants with undisclosed personal characteristics. All models include fixed effects (FE) at the question and period level. We also report the estimated differences in the effect of defaults due to incentives / cognitive load. Significance coding: *** 0.1%, ** 1%, * 5%.*

Model	(1)	(2)
Dep. Variable	Follow default-favored answer	
Question Domain	SG	TTO
Incentives	0.186	0.051
	0.213	0.187
Success probability medical treatment	-1.930 ***	-
	0.234	-
Incentives × Success probability medical treatment	-0.159	-
	0.323	-
Years in full health medical treatment	-	-0.201 ***
	-	0.023
Incentives × Years in full health medical treatment	-	-0.014
	-	0.032
Age	0.005	0.011 *
	0.005	0.005
Male	-0.165	0.286 **
	0.105	0.101
Income	0.033	0.006
	0.020	0.101
Constant	0.992 ***	0.278
	0.266	0.257
Number of participants	420	420
Number of observations	1,260	1,260
Question FE	No	No
Period FE	Yes	Yes
Demographic controls	Reported	Reproted
Participant clustered std. errors	Yes	Yes

Table A.5: Probit regression results for sensitivity analysis SG and TTO. Reported are raw probit coefficients with clustered standard errors at participant level (reported below coefficient estimates in footnote size). The sample is restricted to non-default treatments excluding participants with undisclosed personal characteristics. The dependent variable in all models takes the value 1 if a participant chooses the status quo (i.e. default-favored answer). We include a dummy for the 'Incentives' treatment, variables that capture the success probabilities and years in full health of the medical treatment, and interaction terms between 'Incentives' and the latter. 'Age' measures participants' age in years, 'Male' is a dummy with value 1 if participant is male, and 'Income' represents the 9 different income categories for household income (ascending order, treated as continuous variable). Model (1) and (2) were used to calculate the average marginal effects plotted in ***Figure 7*** *of the main text. All models include fixed effects (FE) at at period level. Significance coding: *** 0.1%, ** 1%, * 5%.*

	Relative frequency of choosing default-favored answer					
	Subjective well-being		Health		Over-claiming	
Treatment	Positive	Negative	SG	TTO	Existing	Non-Existing
Control	32.0%	39.4%	62.2%	43.8%	60.9%	33.8%
[95% confidence interval]	[26.7%, 37.3.%]	[35.3%, 43.6%]	[57.2%, 67.3%]	[38.6%, 48.9%]	[58.1%, 63.7%]	[29.7%, 37.8%]
Control without prediction	36.0%	36.5%	59.6%	43.1%	64.0%	31.0%
[95% confidence interval]	[30.7%, 41.3%]	[32.3%, 40.7%]	[54.8%, 64.3%]	[38.1%, 48.1%]	[61.1%, 67.0%]	[27.3%, 34.7%]
df, *cohen-d*	434, 0.10	434, 0.09	434, 0.07	434, 0.02	434, 0.14	434, 0.10
t-stat, p-value	1.057, 0.29	0.965, 0.34	0.761, 0.45	0.179, 0.86	1.48, 0.13	0.998, 0.32
Difference singinifcant at 5%	No	No	No	No	No	No
Defaults	45.0%	51.7%	64.3%	52.9%	68.8%	39.0%
[95% confidence interval]	[39.4%, 50.6%]	[47.0%, 56.4%]	[59.3%, 69.2%]	[47.5%, 58.2%]	[65.8%, 71.8%]	[34.2%, 43.9%]
Defaults without prediction	41.9%	43.1%	64.8%	51.2%	66.4%	35.9%
[95% confidence interval]	[36.1%, 47.6%]	[38.5%, 47.6%]	[60.1%, 69.4%]	[46.1%, 56.3%]	[63.2%, 69.5%]	[31.2%, 40.6%]
df, *cohen-d*	417, 0.07	417, 0.25	417, 0.01	417, 0.04	417, 0.11	417, 0.09
t-stat, p-value	0.770, 0.44	2.610, 0.01	0.135, 0.89	0.443, 0.66	1.099, 0.27	0.921, 0.36
Difference singinifcant at 5%	No	Yes	No	No	No	No
If defaults, then on	NO	YES	H. State A	H. State A	YES	YES

Table A.6: Relative choice frequencies for default-favored answers across treatments and question domains with 95% confidence intervals. Treatments with and without prediction tasks are contrasted. Differences were assessed via two-sided, independent samples t-tests. Reported are the degrees of freedom (df), cohen-d effect size measures, the t-statistics of the tests (t-stat), and p-values.

Model	(1)	(2)
Dep. Variable	ln(response time)	
Data	Default favored answers	
	rejected	chosen
Winsorization	1%	1%
Incentives	0.025	0.029
	0.034	0.037
Defaults	0.150 ***	-0.144 ***
	0.035	0.039
Defaults & Incentives	0.220 ***	-0.108 **
	0.035	0.037
SG	0.822 ***	0.782 ***
	0.029	0.030
TTO	0.562 ***	0.486 ***
	0.029	0.036
Recognition	0.066 ***	-0.001
	0.017	0.019
Constant	2.318 ***	2.532 ***
	0.064	0.069
Difference in estimated coefficients		
Defaults & Incentives vs	0.070 *	0.036
Defaults	0.035	0.039
Number of participants	864	864
Number of observations	10,509	11,091
Period FE	Yes	Yes
Demographic controls	Yes	Yes
Participant clustered std. errors	Yes	Yes

*Table A.7: Panel generalized-least-squares regression results with robust standard errors reported below the coefficient estimates in footnote size. Model (1) uses questions for which participants rejected the default answer. Model (2) uses those for which participants choose the default answer. Both models include demographic controls (age, income, gender) and period (question order) fixed-effects. Significance coding: *** 0.1%, ** 1%, * 5%.*

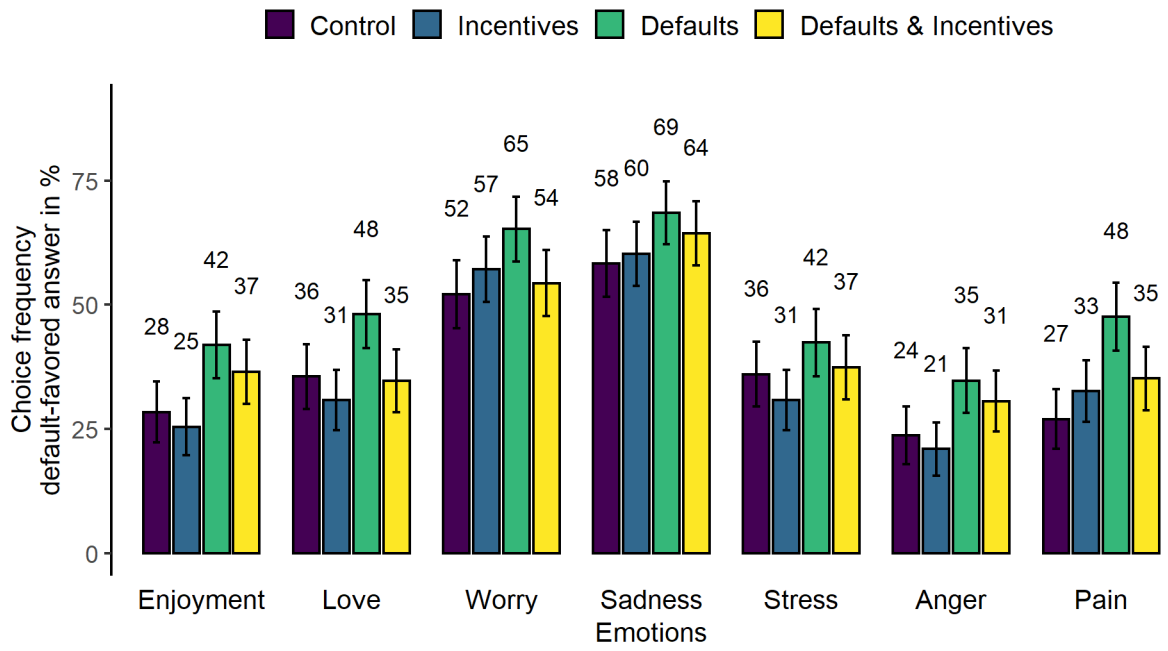

Figure A.1: The proportion of 'yes' responses across treatments and emotions for subjective well-being. For positive emotions the default-favored option was 'NO', for negative emotions it was 'YES'.

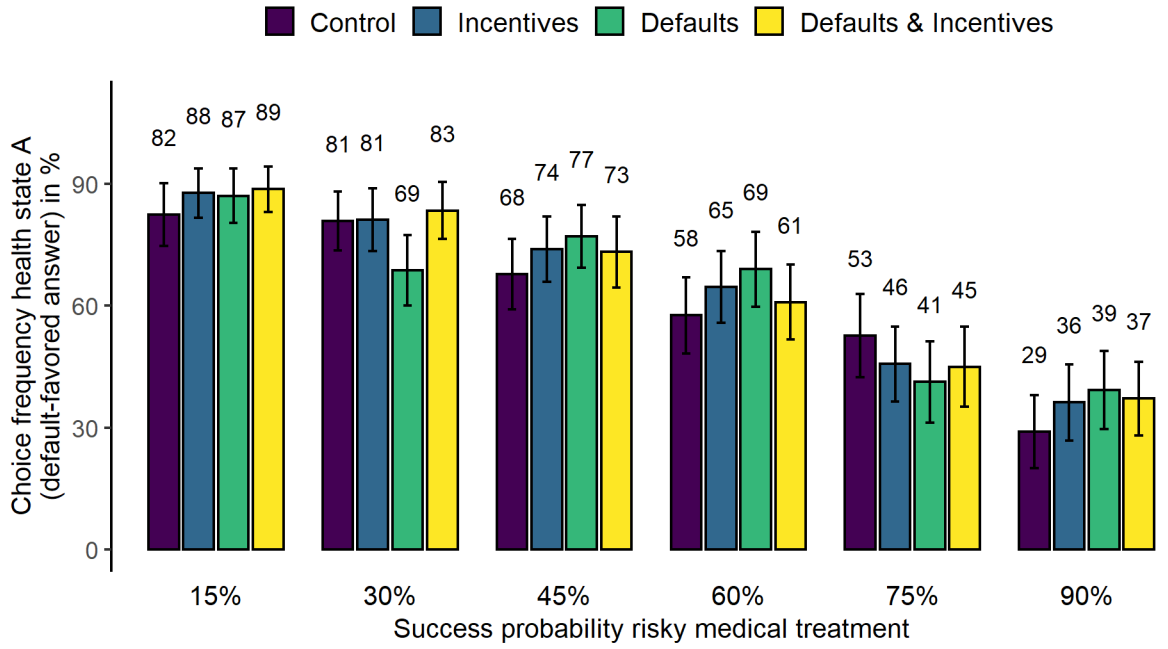

Figure A.2: The proportion of choices for health state A (the default-favored option) across treatments and success probabilities in the standard gamble.

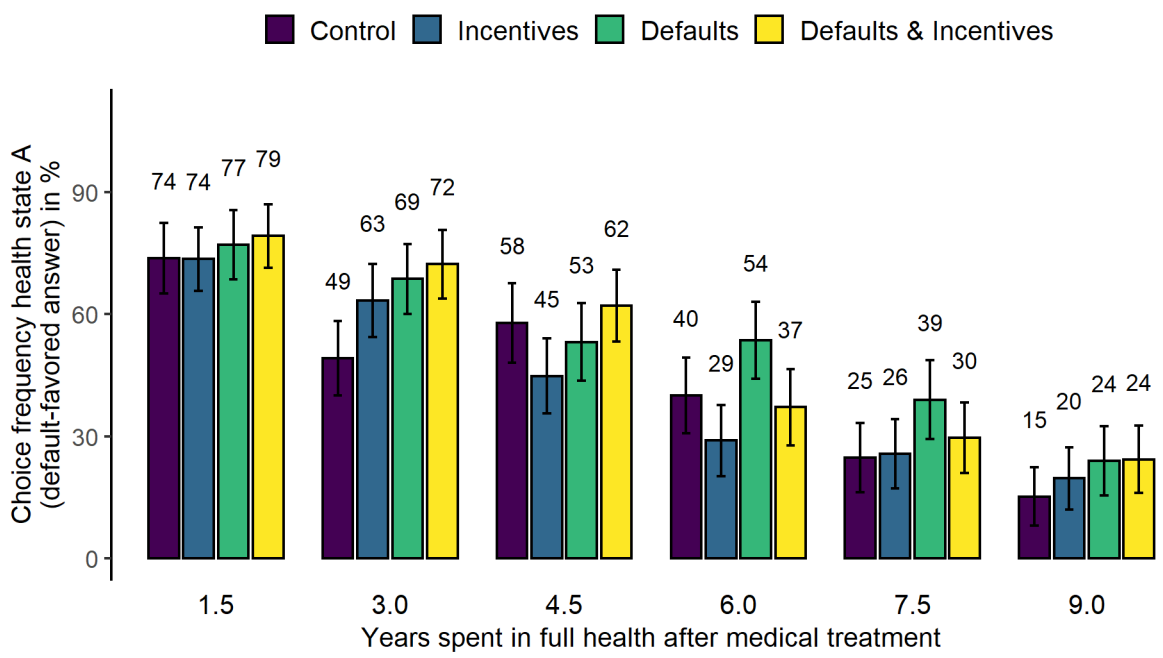

Figure A.3: The proportion of choices for health state A (the default-favored option) across treatments and years spent in full health in the time trade-off.

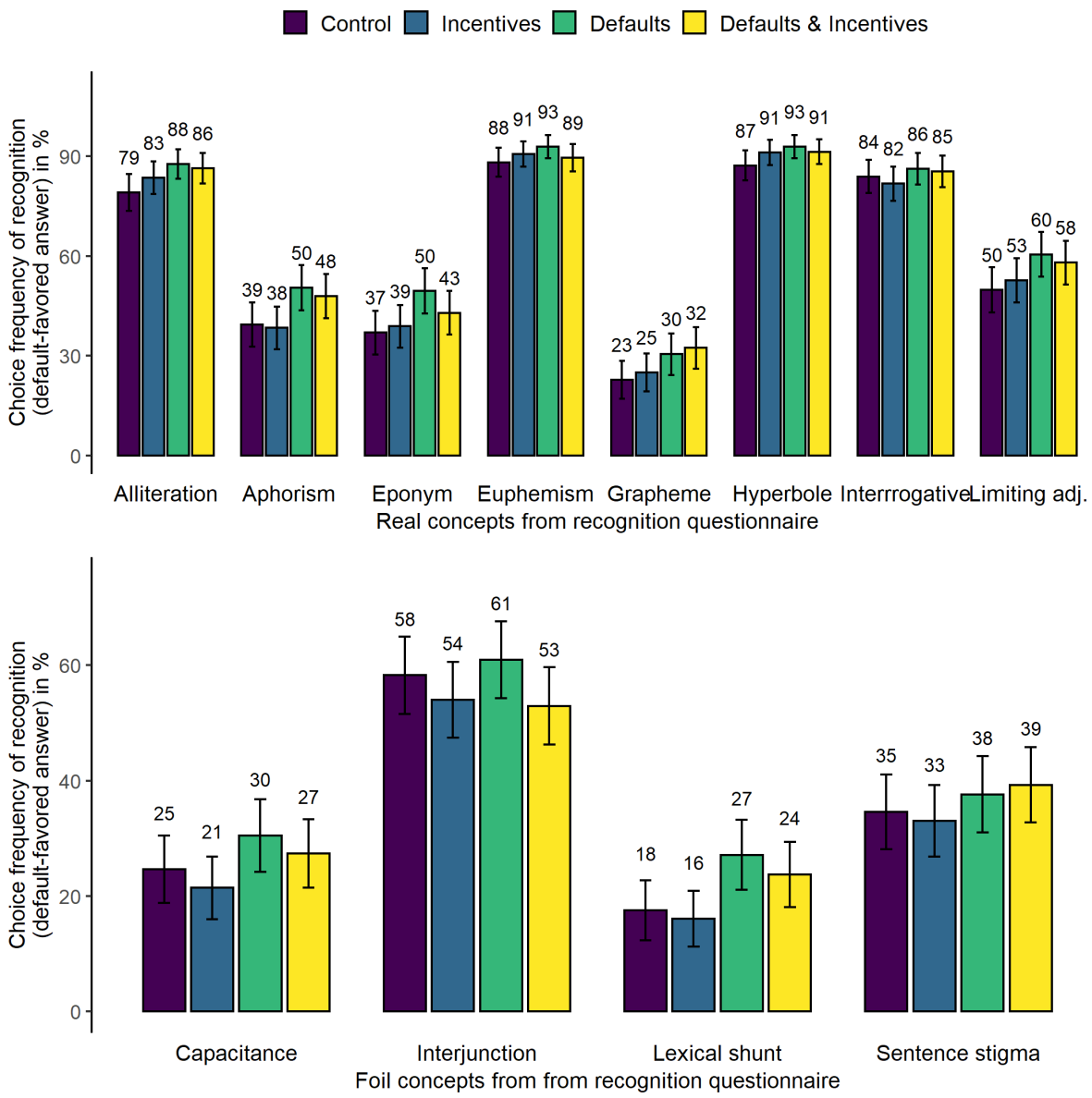

Figure A.4: Recognition rates across treatments and concepts from the recognition questionnaire. The default answer was 'yes'. Top figure: existing concepts, bottom figure: non-existing concepts.

Age	Contrast	Std. Err.	t-stat	p-value
Comparison				
Control vs Incentives	-0.44	1.05	-0.42	0.98
Control vs Defaults	-0.82	1.07	-0.76	0.87
Control vs Defaults & Incentives	-0.44	1.06	-0.42	0.98
Incentives vs Defaults	-1.25	1.05	-1.19	0.63
Incentives vs Defaults & Incentives	-0.88	1.04	-0.85	0.83
Defaults vs Defaults & Incentives	-0.37	1.06	-0.35	0.99

Table A.8: Pairwise mean comparisons for the variable age using Tukey's honestly significant difference test to correct for multiple comparisons.

Time taken in secs	Contrast	Std. Err.	t-stat	p-value
Comparison				
Control vs Incentives	-22.06	31.51	-0.70	0.90
Control vs Defaults	-3.83	32.01	-0.12	1.00
Control vs Defaults & Incentives	61.26	31.68	1.93	0.22
Incentives vs Defaults	-25.89	31.55	-0.82	0.85
Incentives vs Defaults & Incentives	39.21	31.21	1.26	0.59
Defaults vs Defaults & Incentives	-65.09	31.72	-2.05	0.17

Table A.9: Pairwise mean comparisons for seconds spend on taking the survey using Tukey's honestly significant difference test to correct for multiple comparisons.

7. Appendix B: Survey questions and survey design

This sub-section contains a detailed transcript of the survey instructions. For repeating elements, e.g. emotions, we report the instructions for one representative element. Treatment specific text passages are marked correspondingly. Prolific is a UK-based company and all payments are processed and communicated in British Pounds. We used the historical ex-change rate to convert pound payments to US Dollars as reported in the main text. Sample screenshots from the decision screens are provided in **Figure 2** and **Figure 3**.

Introduction.

(All) Thank you for participating in this survey. The first part of this survey asks questions about your well-being and health. The second part asks whether or not you recognize various items. For each question we will also ask you to estimate the percentage of participants giving specific answers to the question.

(Control and *Defaults)* On top of your fixed earnings of £1, you will earn a bonus payment of £1 for completing this survey.

(Incentives and *Defaults & Incentives)* For each completed question, you will earn points based on that answer's "Truth Score". Truth Scoring was invented by an MIT professor and published in the journal SCIENCE, which is one of the most prestigious scientific journals. Truth Scoring rewards you for answering truthfully. Even though only you know if you have answered truthfully, the Truth Scoring algorithm ensures that people who tell the truth score higher overall. On top of your fixed earnings of £1, you will earn a bonus payment depending on your Truth Score. The higher your Truth Score is, the higher your bonus. You maximize your bonus if you answer every question truthfully. So it is in your best interest to consider each question thoroughly and to answer each question truthfully. Your bonus payment can be as high as £2 (on average it is £1).

(All) Please read all questions carefully and follow the on-screen instructions. Answer honestly and take care to avoid mistakes. Completing the survey will take about 12 minutes. Please click on Next to proceed.

Subjective well-being.

(All) You will now be presented with a series of questions about your well-being. Please click on Next to proceed.

(Incentives and *Defaults & Incentives)* Please keep in mind that you will earn a bonus depending on your Truth Score. It is in your best interest to consider each question thoroughly and to answer each question truthfully.

(All, new screen) Drag and drop YES and NO buttons to the *enjoyment* box to indicate your answer. Did you experience *enjoyment* during a lot of the day yesterday?

(All, new screen) The previous questions asked "Did you experience *enjoyment* during a lot of the day yesterday?" Please estimate how many out of 100 participants answered YES on this question.

Standard gamble and Time Trade-Off.

(All) You will now be presented with a series of questions about health scenarios. The health scenarios are described below. They differ on several dimensions relating to mobility, self-care, ability to perform your usual activities, pain, and anxiety/depression. Health state A: - Moderate problems in walking about, - Moderate problems with self-care activities (e.g. washing or dressing), - Unable to perform usual activities (e.g. work, study, family or leisure activities), - Severe pain or discomfort, - Moderately anxious or depressed. Full health: - No problems in walking about, - No problems with self-care activities (e.g. washing or dressing), - No problems with performing usual activities (e.g. work, study, family or leisure activities), - No pain or discomfort, - Not anxious or depressed. Please click Next to proceed.

(Incentives and *Defaults & Incentives)* Please keep in mind that you will earn a bonus depending on your Truth Score. It is in your best interest to consider each question thoroughly and to answer each question truthfully.

Standard Gamble.

(All, new screen) Imagine yourself living the rest of your life in Health state A. You can choose to take a risky medical treatment. Drag and drop the scenario buttons to the 'I prefer' box to indicate your answer. Which scenario do you prefer? Scenario 1: Living the rest of your life in Health state A. Scenario 2: Taking a risky medical treatment with two possible outcomes. With probability 0.15 you live in full health for the rest of your life. With probability 0.85 you die within one week.

(All, new screen) The previous questions asked you to choose between living in health state A and the risky medical treatment in which you live in full health with a probability of 0.15. Please estimate how many out of 100 participants choose to live in health state A.

Time trade-off.

(All, new screen) Imagine yourself living for 10 years in Health state A. You can choose to take a medical treatment. Drag and drop the scenario buttons to the 'I prefer' box to indicate your answer. Which scenario do you prefer? Scenario 1: Living in Health state A for 10 years after which you die. Scenario 2: Taking a medical treatment and living in Full health for 1.5 years after which you die.

(All, new screen) The previous questions asked you to choose between living in health state A for 10 years and living in full health for 1.5 years. Please estimate how many out of 100 participants choose to live in health state A for 10 years.

Recognition questionnaire.

(All) You will now be presented with several items. For each item, please indicate whether you recognize it, that is, whether or not you know it in the specified context. Only you know whether or not you recognize an item. Please note that some items may not exist in the specified context. Please click on Next to proceed.

(Incentives and *Defaults & Incentives)* Please keep in mind that you will earn a bonus depending on your Truth Score. It is in your best interest to consider each question thoroughly and to answer each question truthfully.

(All, new screen) Drag and drop YES and NO buttons to the *Alliteration* box to indicate your answer. Please indicate whether or not you recognize *Alliteration* as a concept from the language arts.

(All, new screen) The previous questions asked whether or not you recognized *Alliteration* as a concept from the language arts. Please estimate how many out of 100 participants recognized *Alliteration*.