How to Improve Response Consistency in Discrete Choice Experiments? An Induced Values Investigation

Nicolas Jacquemet* Stéphane Luchini† Jason F. Shogren‡ Verity Watson§

March 2016

Abstract

Discrete choice experiment surveys are used to elicit preferences for multi-attribute goods and services. While the implicit presumption in the survey is that people choose the alternative that maximizes their satisfaction, it remains an open question whether answers collected in DCE surveys can generate reliable welfare estimates for policy analysis. Herein we step back into the lab to design an experimenter-controlled, pre-assigned induced value experiment to explore truth-telling within a DCE framework. Our baseline treatment finds weak truth-telling. This insincere behavior is neither improved by (i) helping respondents reduce their cognitive burden nor by (ii) asking them to sign an oath to "faithfully and conscientiously fulfill" their duties. In contrast, our results suggest we can significantly improve the reliability of elicited preferences by asking subjects to sign a solemn oath that commits them to truth-telling. This finding suggests that hypothetical bias arises not from task complexity or lack of realism, but rather the lack of commitment to truthfulness.

Keywords: Stated preference, External validity, Welfare, Oath, Truth-telling, Survey methods.

JEL codes: D1, D6.

Significance: People have preferences over public goods and novel private goods not yet priced by the market. These goods have many attributes that people like, and the economist's challenge is to put an accurate monetary value on these preferences. The discrete choice experiment (DCE) is one popular method to do this. The open question is whether one can use DCE answers to measure welfare for policy analysis, given people are responding to a survey rather than facing their own budget constraint and spending their own money. Stepping back into an experimenter-controlled lab setting, we find insincere bidding is common in the DCE. We then show that the reliability of elicited preferences in DCE drastically improves once people sign a solemn oath that commits them to truth-telling.

Introduction

Stated preference methods use hypothetical surveys like Discrete Choice Experiments (DCE) to elicit preferences for multi-attribute public goods and services, like environmental quality and health care.[1] A typical DCE choice

*Paris School of Economics and University Paris 1 Panthéon-Sorbonne (CES).

†Université d'Aix-Marseille (Aix-Marseille School of Economics, CNRS & EHESS.

‡Department of Economics and Finance, University of Wyoming.

§HERU, University of Aberdeen.

[1]Fifteen years ago, R. Carson [1] estimated that at least 2000 stated choice studies had been run. Examples abound, including the most famous study to measure environmental damages from the 1988 Exxon Valdez oil spill. Other examples include food safety

Token	**Size**			**Colour**			**Shape**			**Cost**		
Attributes	Small	Medium	Large	Red	Yellow	Blue	Circle	Triangle	Square			
£	0.50	2.50	4.00	1.00	1.50	2.00	1.50	3.00	6.00	2.00	3.00	4.00

Table 1: Subjects' induced values for all treatments

task in these surveys asks a person to choose between a set of multi-attribute bundles of goods. While gaining acceptance as a mainstream technique to inform public policy and legal battles, the question remains how reliable answers to hypothetical surveys are as a measure of reality. This external validity issue, due to the so-called hypothetical bias problem, is at the core of the non-market valuation literature [see, for example, 2].

In a recent PNAS paper, [3] substantiate the use of DCEs to guide policy based on their predictive validity – a well-designed DCE survey can do a reasonable job of predicting actual behavior, whether or not it is rational. For that purpose, the careful selection of the sample and questions matter: "*to maximize external validity about real-world causal effects, survey samples need to be carefully chosen to match the target population and survey experimental designs need to be carefully crafted to motivate respondents to seriously engage with hypothetical choice tasks to mimic the incentives that they face when making the same choices in the real world.*" (p.2400). Predictive validity however leaves open the question of the ability of DCE to measure and convert behavior into the money metrics commonly used in cost-benefit analysis (CBA). CBA are a very common application of DCE in, say environmental or health policy. The money metrics are essential to estimate the gains/losses in welfare of any proposed change. To that end, one needs to know what motivates a persons choices when answering DCE questions: is this behavior best captured by rational choice or by bounded rationality? If DCE survey responses are boundedly rational, such behavior needs to be better understood so they can be removed effectively from welfare analysis [e.g., see 4], either *ex ante* through better survey design or *ex post* through calibration of real and hypothetical behavior. But understanding the behavior at work in a DCE requires more control than can be delivered in the field. In this paper, we step back into the lab to gain this control. We carry out an induced valuation experiment to test behavior within a DCE setting – our goal is not to predict behavior assuming rational choice exists, but rather to examine actual behavior relative to a theoretical benchmark of rationality. We use experimental tools to show that problems arise in DCE – not because of bounded rationality but rather because respondents are not truthful. We then show how using a solemn oath to tell the truth in the DCE can resolve this question of strategic dishonesty.

Induced value discrete choice experiment

Our induced value experimental design recreates the salient features of a stated choice study in an experimental economics laboratory setting. The baseline experiment is the wide hypothetical treatment carried out in [5]. We induce preferences for a multi-attribute good that we call a token. Subjects may buy tokens during the experiment at the announced cost, which they will then sell back to us at the end of the experiment. Subject's preferences for the tokens are induced by announcing that the amount we will pay for a token depends on its attributes and levels. A token has four attributes and each attribute has three levels: colour (red, yellow, blue); shape (circle,

from pathogens, biodiversity and endangered species protection, the reintroduction of Gray Wolves to Yellowstone National Park and salmon restoration in New England. In health care, clinical trials measure the health improvement using a five attribute Quality Adjusted Life Year (QALY) measure known as the EQ-5D. The new value set for the EQ-5D is based on responses to a stated choice survey. In the United Kingdom (UK), the HM Treasury Green Book suggests stated choice studies as a method to obtain evidence to use in cost-benefit analysis of projects, such as infrastructure investments.

triangle, square); size (small, medium, large); and cost. The levels of each of the attributes is associated with a monetary value, as shown in Table 1. The sum of the attribute values determines the value of the token, i.e. the amount the monitor will pay to buy it back from the subject. This replicates the linear additive utility typically assumed in DCE studies.

Subjects are asked to complete nine choice tasks on a computer with one choice task per screen. In each task they are offered two tokens, and can choose either to buy one of the two tokens to sell back to us later, or to buy no token at all.[2] The best choice for a subject is to buy the token with the highest profit in each choice set, i.e. the token in which the difference between the value (sum of attribute levels) and cost is greatest. The monetary value of attributes induces subject's preferences over the attributes in the choice sets – profits are the lab counterfactual of individual preferences in real life. This induced value design allows us to assess whether subjects make the *best choice* for them, i.e. make choices that maximize their profit/experienced utility. This cannot be done in a homegrown preferences DCE because the preferences underlying elicited choices are respondent's private information. It also allows us to assess how best choices vary across experimental treatments. Our seven experimental treatments are detailed below. We focus on the effect of the treatments on three outcome variables. First, we assess the reliability of the stated choice method at the aggregate level based on the overall proportion of best choices. Second, at the subject level, we look at the total number of best choices made by a subject. Third, we record the time taken for a subject to make each choice to the nearest second. We use response time as a proxy for cognitive reasoning in the task, assuming that subjects who respond more slowly engage in more cognitive reasoning.[3]

Experimental treatments

Table 2 summarizes our baseline experiment and six treatments. All treatments are implemented between subjects – i.e., the same person participates in only one of the seven experiments. Consider each in turn.[4]

Experiment 1 (*baseline*): Subjects' choices are hypothetical as in a DCE survey. Subjects are paid £12 for taking part in the experiment irrespective of the choices they make (tokens they buy). The experiment instructions use subjective language by asking individuals, e.g., to "*put yourself in a situation where your account balance at the end of the experiment would depend on the choice you made...*" [11]. We take experiment 1 as our baseline treatment to which we compare all other treatments. The choice tasks and tokens are identical for all experimental treatments.

Experiment 2 (*calculator*): We provide subjects with a computerized calculator to help them make the calculations. Several studies suggest that the DCE tasks are too complicated for respondents, and, as a consequence, respondents might not choose what is best for them. In particular, choices are more complicated when the multi-attribute goods included in the choice set are similar [12, 13], when the goods are described by many attributes

[2]The tokens included in the choice tasks were chosen using a fractional factorial design.

[3]This relationship between cognitive effort and response time (RT) is confirmed by recent empirical evidence [6]: the more obvious a choice is, the lower is RT [7]. It is worth highlighting we do not aim to perform reverse inference – i.e. identify the mental process at work in decision making (e.g., intuitive vs deliberative reasoning) based on RT variations, which requires that unobserved components of decision-making are controlled for [8].

[4]Subjects who take part in our experiments are students at the University of Aberdeen, who are recruited to the experiments using Exlab and ORSEE software [9]. All subjects received a consent form, experiment instructions, and payment form before taking part in the experiment. Before the experiment started, the subjects read and signed the consent form and this was collected by the experimenter, then the experimenter read aloud the experiment instructions to the group and answered questions. The experiment was programmed and conducted with the software z-Tree [10]. The computer based experiment means that the order of choice sets was randomized allowing us to separate choice sets effects from decision round effects.

Experimental treatments	Design	Hyp. choice	Calculator	Oath
Experiment 1. *Baseline*	Each subject completes 9 choice tasks: (1 choice per screen). Task: Offered two tokens. He or she chooses to (1) buy one token to sell back, or (2) buy no token	Yes	No	No
Experiment 2. *Calculator*	Each subject has access to the Microsoft Windows ^{TM} calculator	Yes	Yes	No
Experiment 3. *Paid*	We pay each subject based on his or her choices. Payoffs are determined by randomly selecting 1 round out of 9	No	No	No
Experiment 4. *Calc. + paid*	Combines experiments 2 and 3	No	Yes	No
Experiment 5. *Oath on truth*	"I, ..., the undersigned do solemnly swear that during the whole experiment, I will tell the truth and always provide honest answers"	Yes	No	Yes
Experiment 6. *Oath on task*	"I, ..., the undersigned do solemnly swear that during the entire experiment, I will faithfully and conscientiously fulfill the tasks that I am asked to complete to the best of my skill and knowledge"	Yes	No	Yes
Experiment 7. *Oath on duty*	"I, ..., the undersigned do solemnly swear that during the whole experiment, I will faithfully and conscientiously fulfill my duties to the best of my skill and knowledge"	Yes	No	Yes

Table 2: Summary of baseline and treatments

[13, 14], when the choices sets include many alternatives [14], or when individuals are asked to answer many choice tasks [14]. Although our induced value DCE tasks may seem to involve basic mathematics (addition and subtraction) not all subjects may be able to complete this task, therefore we provide the help of a calculator.

Experiment 2 replicates experiment 1, but with a button added to each choice set screen, by clicking on this button subjects can access the Microsoft windows^{TM}calculator. Subjects' calculator use is recorded throughout the experiment. In the experiment instructions, subjects are told how to access and use the calculator. Otherwise, experiment 2 is identical to experiment 1.

Experiment 3 (*paid*): We pay subjects based on the choices that they make in the experiment. In experiments 1 and 2, choices are hypothetical and do not affect how much subjects earn in the experiment. Critics of stated preference methods, and survey methods in general, question peoples' motivation to choose the best for them when answers are hypothetical [see, e.g., 15]. Subjects' intrinsic motivation alone may not be enough to engage them in making the necessary cognitive effort to solve the task [see, e.g., 16, for a discussion of this issue in economic experiments]. In experiment 3, we replicate experiment 1, but the choices that subjects make affect their earnings from the experiment.

To that end, subjects receive a £4 show-up fee with which to buy a token (all tokens offered for sale in the experiment cost less than £4). Subjects face the same nine choices as in experiment 1. At the end of the experiment, one of the subject's nine choice tasks is selected at random to be binding: the actual choices to that choice task are used to compute the subject's earnings from the experiment. Randomly selecting the round as the payoff trial prevents subjects' previous choices from influencing the money that they have to spend in each round. Experiment 3 is identical to experiment 1 and the experimental instructions are identical to those for experiment 1 except that they do not use subjective language.

Experiment 4 (*paid+calculator*): We combine monetary incentives with a calculator. Our hypothesis is that when choices affect earnings this fosters cognitive reasoning and may encourage subjects to use the calculator more. This would, in turn, lead to a higher proportion of best choices. In economic terms, monetary incentives and the calculator could act as complements to improve decision making.

Experiment 5 (*oath on truth-telling*): We implement a non-priced commitment device – a truth-telling oath – in an hypothetical setting similar to experiment 1. The truth-telling oath – similar to that taken by witnesses before giving evidence in a court of law – has been found to improve individuals' behavior in both induced and homegrown value auctions [17], and [18] have shown in a DCE survey that a truth-telling oath can reduce the gap between hypothetical choices and real economic choices. The effect of the truth-telling oath can be explained as follows. Individuals' survey responses are not binding. In the absence of a direct link between individuals and their declaration, individuals may lack the necessary commitment to provide reliable answers. Taking an oath restores this link by committing people to truth-telling [see 17, 19, for more details].

The truth-telling oath procedure in experiment 5 follows that of [17, 19]. Subjects are presented with the oath at a private desk upon entry into the lab, but after completing the consent form. The form is entitled "Solemn oath" and contains an unique sentence with a single prescription that reads "I, ..., the undersigned do solemnly swear that during the whole experiment, I will **tell the truth and always provide honest answers**" (the oath form is presented in Appendix A). Subjects are told that signing the form is voluntary and that neither their participation in the experiment nor their earnings depend on signing [see 19, who explain this choice by borrowing insights from the social psychology of commitment].

Experiment 6 (*oath on task*): We implement a modified oath that targets cognitive effort. This allows us to test whether the truth-telling oath works by fostering cognitive reasoning. Experiment 6 replicates experiment 5, but with a modified oath form that explicitly targets cognitive effort without referring to truth-telling behavior. The oath form now reads "I, ..., the undersigned do solemnly swear that during the entire experiment, I will **faithfully and conscientiously fulfil the tasks that I am asked to complete to the best of my skill and knowledge**". Otherwise, the oath form and the oath procedure are identical to that of experiment 5.

Experiment 7 (*oath on duty*): We implement a second modified oath that targets cognitive effort with a moral component. The oath in experiment 5 had a moral component (truth-telling) and the oath in experiment 6 targeted cognitive effort. In experiment 7, we again adapt a real world oath, in this case one that targets effort to perform one's assigned task with the moral reminders that one would encounter in the field if taking an oath before beginning the duties of a public office: the *oath of office*. In experiment 7, we carry out the same oath procedure as experiments 5 and 6, using an oath form that now reads "I, ..., the undersigned do solemnly swear that during the whole experiment, I will **faithfully and conscientiously fulfill my duties to the best of my skill and knowledge**". Providing ethical standards to people, which the truth-telling oath and the oath of office do, has been shown to have significant effect on behavior [20].

All oath procedures in experiments 5, 6 and 7 were carried out by the same person for all subjects – she also ran experiments 1 to 4.

Results

Table 3 shows the subjects' profit from selling each token in the nine choice sets (A - I) back to us, the profit difference between the two tokens, and the proportion of subjects who chose the best (highest profit) token in each of the experiments.

Choice	Value Token 1	Value Token 2	Profit	**Exp. 1** *Baseline* (n=47)	**Exp. 2** *Calc.* (n=47)	**Exp. 3** *Paid* (n=54)	**Exp. 4** *Paid+Calc.* (n=47)	**Exp. 5** *Truth* (n=44)	**Exp. 6** *Task* (n=44)	**Exp. 7** *Office* (n=44)
A	5.5	6.5	1.0	14.9	4.3	5.0	5.2	59.1	10.8	0.0
B	2.5	9.5	7.0	38.3	36.9	33.3	30.8	86.4	35.1	24.3
C	3.5	8.0	4.5	14.9	13.0	27.7	10.3	84.1	18.9	8.1
D	-0.5	7.0	7.5	76.5	78.3	85.2	87.1	90.9	89.2	91.9
E	8.0	3.0	5.0	72.3	80.4	74.1	84.6	77.3	83.8	94.6
F	4.5	3.0	1.5	72.3	80.4	74.1	89.7	65.9	75.7	75.7
G	6.0	4.0	2.0	74.4	84.8	81.5	94.9	81.8	86.5	91.9
H	3.0	0.5	2.5	68.1	82.6	79.6	87.2	77.3	83.8	75.7
I	8.0	1.0	7.0	74.4	93.5	74.1	94.9	79.5	89.2	91.9
Overall				**56.3**	**61.6**	**59.9**	**64.9**	**78.3**	**63.7**	**61.6**

Table 3: Token values, profit and proportion (in %) of correct choices by choice set and treatment

In **experiment 1** (*baseline*), there are two main results at the aggregate level. First, only a little over half of the choices are the best (56.3%) with no significant difference across rounds (see Appendix B). Second, the proportion of choices that maximize profit varies greatly between choice sets, from 14.9% in choice sets A and C to 76.5% in choice set D.

Choice sets A, B and C have the worst choices, and share the feature that while token 1 is profitable, token 2 is more profitable. Of the three choice sets B has the highest proportion of best choices (38.3%) and the smallest profit for token 1. In contrast, token 2 in choice set D has a high profit and token 1 makes subjects a loss (i.e. the value is less than the cost). Overall, the choice sets with the highest proportion of best choices have one token with zero or very small profits.

In Figure 1, we compute, for each subject, the percentage of best choices made in the 9 choice sets and we present its empirical distribution function (EDF). Each bullet in the figure corresponds to a subject. No subject made 100% (or 9) best choices. The highest percentage of best choices observed in the experiment is 77.7%, which corresponds to 7 choices out of 9. Most of the subjects (44.7%) made only 6 best choices in experiment 1. In many choice sets, subjects do not choose what is best for them in our straightforward DCE design.

One explanation for this result may be that subjects make mistakes when making choices. In particular, in choice sets A, B and C in which token 1 is profitable but token 2 is more profitable. If subjects make mistakes, then we should observe significantly shorter response times for choices that are not best because subjects making mistakes engage less in cognitive reasoning [21]. This is not what the results suggest. When all choice sets are considered together, median response time for bad choices is 15 seconds compared to 17 seconds for best choices. For choice sets A, B and C, median response time of best choices is shorter than bad choices (12 seconds compared to 17 seconds). The response time difference is even greater when only choice sets A and C are considered: median response time for best choices is 8 seconds compared to 17 seconds for bad choices. Bootstrap tests indicate that the increase in median response time is statistically significant for choice sets A and C ($p = .052$ and $p = .001$) whereas there is no significant difference in median response time for choice set B ($p = .358$). We observe an opposite pattern for choice sets D-I. Median response time is longer for best choices (17 seconds) than for bad choices (12.5 seconds). Bootstrap tests indicate that the difference is significant for choice sets D ($p = .024$), G ($p = .011$) and I ($p = .028$). Results on median response time are confirmed by Kolmogorov-Smirnov (KS) bootstrap distribution tests (see Appendix C for EDF graphs and KS tests).

From experiment 1 (baseline), we observe: 1/ the proportion of best choices is very low and casts doubt on the use of DCE results for policy decisions. 2/ bad choices take longer than best choices in those choice sets with the

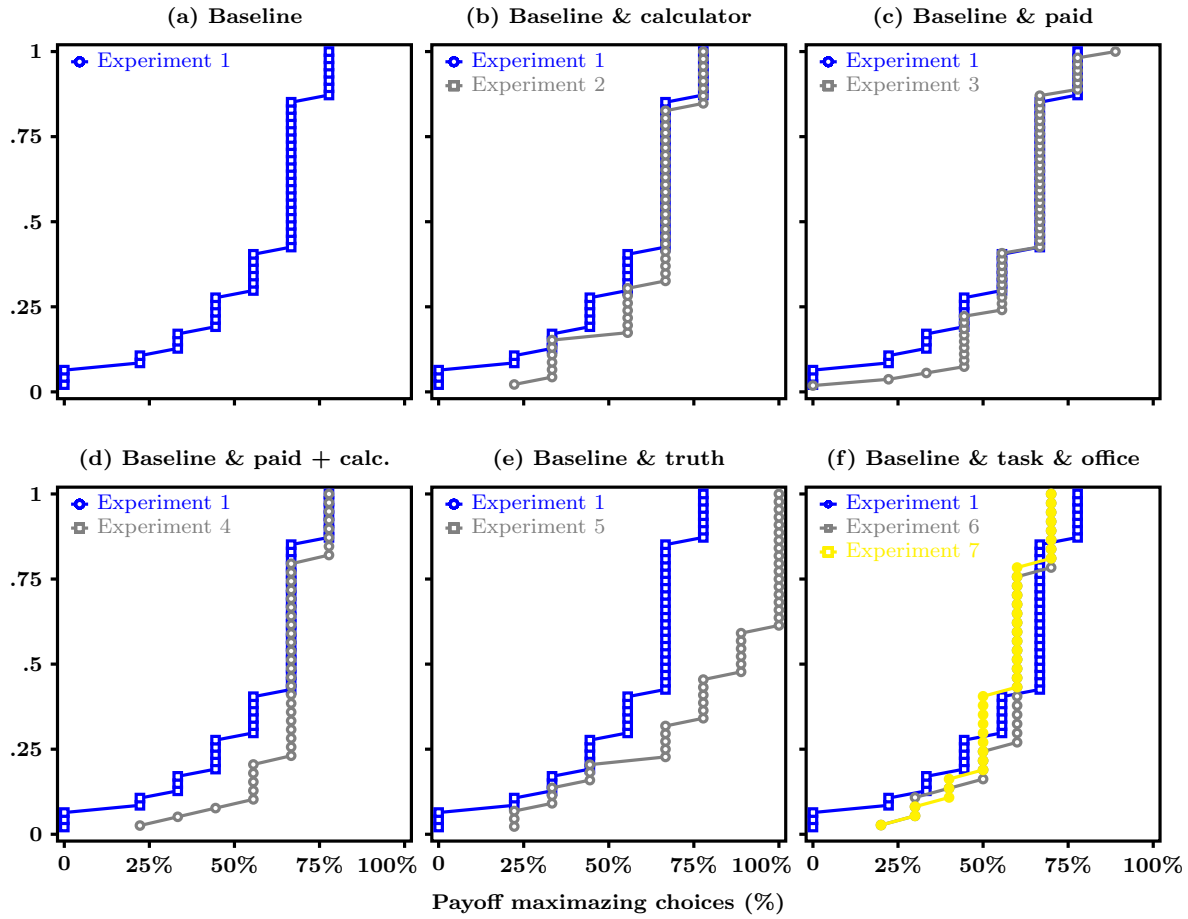


Figure 1: Empirical distribution function (EDF) of proportion of payoff maximizing choices by subject

lowest proportion of best choices. Bad choices are not "mere" mistakes based on intuitive responses rather than cognitive reasoning, but may involve more complex choice processes.

In **experiment 2** (*calculator*) subjects can use the windows calculator to help them make their choices. We record if the calculator is used in a choice task, and how many keyboard entries are made when the calculator is activated. One keyboard entry corresponds to a number, an operator, a decimal mark or a delete key. For instance, a subject who would calculate the value of a small yellow square token would type "`.5 + 1.5 + 6 =`" and this would be counted as 9 keyboard entries.

Figure 2.A presents the proportion of subjects who activated the calculator and the mean number of keyboard entries across rounds and by choice set in experiment 2. The calculator was activated in 24.6% of the choice tasks. Fifty percent of subjects never activated the calculator, 19.5% activated it only once and 13% activated it in every round, the remaining subjects are equally distributed in between. Figure 2.A.a shows that the activation of the calculator is relatively stable across rounds. There is no clear round effect, with only a small increase in activation in rounds 2 and 3 (28.2% and 32.6% respectively and 21.7% in round 1). We do not observe a round effect in number of keyboard entries (Figure 2.A.b), when the calculator was activated (except in round 1). The proportion of activation does not depend on the choice set (Figure 2.A.c and Figure 2.A.d).

For choice sets A, B and C (those choice sets that exhibit a low percentage of best choices in experiment 1) the proportion of activation is 26.1% whereas it is 23.9% for the remaining choice sets D to I. When the calculator

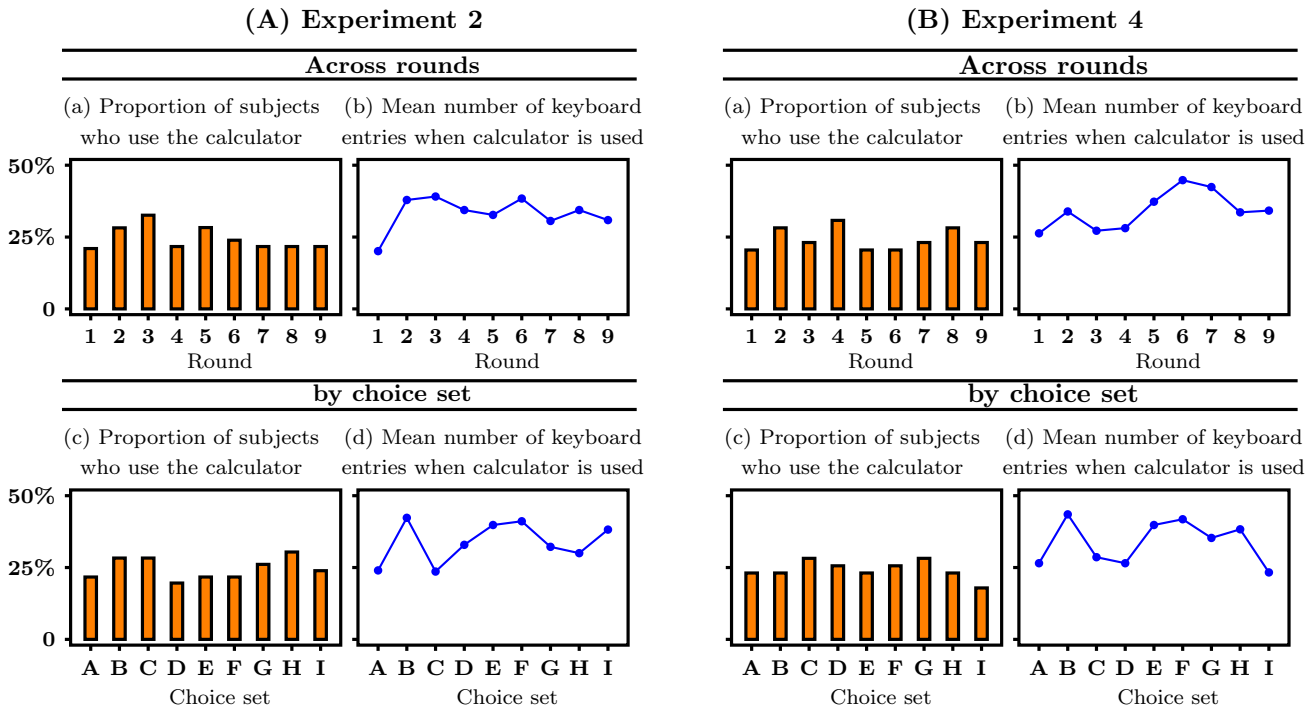


Figure 2: Use of the calculator across rounds and by choice set

was activated, the mean number of keyboard entries was also very similar across choice sets. The two choice sets with the lowest mean number of keyboard entries were also those with the lowest percentage of best choices in experiment 1 (A and C).

Table 3 presents the percentage of best choices in **experiment 2**. Providing a calculator has a small but statistically insignificant effect on the percentage of best choices compared to **experiment 1** (61.6% vs. 56.3%). The two proportions are compared using a two-sided bootstrap test of proportions that allows for within subject correlation, the p-value is $p = .298$. The percentage of best choices for choice sets D-I is 83.3% in experiment 2 compared to 73.0% in experiment 1, but the difference is not significant ($p = .176$). Similarly, the calculator does not significantly change the percentage of best choices in choice sets A, B and C – in these choice sets the share of best choices remains low.

At the subject level, we observe no improvement in the percentage of best choices a subject makes. Figure 1.b presents the EDF of the percentage of best choices by subjects in **experiments 1 and 2**. The EDF in experiment 2 is slightly to the right of the EDF in experiment 1 but first order dominance is not significant ($p = .192$).[5]

We test if subjects who use the calculator make better choices. We find that in choice sets D to I, subjects who use the calculator make best choices 92.5% of the time and subjects who do not use the calculator make best choices 80.5% of the time. In choice sets A, B and C, individuals use the calculator make best choices 16.7% of the time compared to 18.6% for subjects who do not use the the calculator. Pairwise correlation between the number of times the calculator is used by subject and the total number of best choices made is positive but not statistically significant for choice sets D to I: .271 with $p = .128$. Pairwise correlation is negative, equal to -0.140,

[5]We use a bootstrap version of the Kolmogorov-Smirnov test. The advantage of this test as compared to the standard KS test is to allows for ties and small sample size [see 22, 23].

but not significant for choice sets A, B and C ($p = .649$). In this experiment, providing help to improve cognitive reasoning does not improve choices.

In **experiment 3** (*paid*), subjects are paid based on their choices. Table 3 presents the percentage of best choices by choice set. Overall, we observe that 59.9% of choices are best when money is at stake. This is not statistically different from that of **experiment 1** ($p = .607$). The percentage of best choices in choice sets D to I is 78.1%, close to that observed in experiment 1 (73.0%). There are no differences for choice sets A, B and C: 23.5% in experiment 3 and 22.7% in experiment 1. Figure 1.c, that plots the EDF of the percentage of best choices by subjects in experiment 1 and 3, confirms aggregate findings at the subject level: the EDF are very much alike and although the EDF in experiment 3 is slightly to the right of the EDF in experiment 1, first order dominance is not significant ($p = .480$).

Subjects take more time to make their choices in experiment 3. Median total response time (time taken by a subject to answer all nine choice sets) is 197 seconds in experiment 3 compared to 157 seconds in experiment 1, the increase is significant with $p = .050$ (median difference bootstrap test). The EDF of total response time shows that being paid reduces the number of subjects with very quick response times (see figure in Appendix D). A KS bootstrap distribution test indicates that the EDF of response time in experiment 3 first order dominates the EDF of response time in experiment 1 ($p < .025$). Longer response times seem to be significantly associated with better choices: pairwise correlation is .349 with $p = .010$. Subjects who take longer make better choices.

In **experiment 4** (*paid+calculator*) subjects are paid based on their choices and can use a calculator to help them make choices. Figure 2.B presents calculator use in experiment 4, across rounds and by choice set. Being paid based on choices made does not increase calculator use compared to experiment 2. The calculator was activated in 24.2% of the choice tasks in experiment 4 compared to 24.6% of the choices tasks in experiment 2. Again, nearly fifty percent of subjects (48.7%) never use the calculator and 12.8% use it in every choice set (13% in experiment 2). As in experiment 2, there is little evidence of round or choice sets effects. Among those subjects who actually use the calculator, we do not find differences in the number of keyboard entries between experiment 4 and experiment 2. Overall, mean number of keyboard entries is 33.9 in experiment 4 and 33.6 in experiment 2.

Table 3 presents the percentage of best choices by choice set. Being paid in combination with the calculator leads to a small, but significant, increase in the percentage of best choices: 64.9% compared to 56.3% with $p = .037$. This improvement comes from better decisions in choice sets D to I: 89.8% of best choices in experiment 4 and 73.0% in experiment 1 ($p = .005$.). There is no significant change in the percentage of best choices in the most problematic choice sets A, B and C: 15.4% in experiment 4 and 22.7% in experiment 1 ($p = .328$).

Figure 1.d presents the EDF of the percentage of best choices by individual in experiment 1 and experiment 4. The EDF in experiment 4 is shifted to the right: there are fewer subjects with a low number of best choices because of the improvements for choice sets D to I. The upper part of the EDF remains similar to that of experiment 1 because of the unchanged proportions for choice sets A, B and C. The EDF in experiment 4 first order dominates the EDF in experiment 4 ($p = 0.015$).

In choice sets D to I, there is no difference in payoff maximizing choices between those who use the calculator (92.9%) and those who do not (88.8%). These proportions were respectively 92.5% and 80.5% in experiment 2. In choice sets A, B and C, 17.2% of choices made with the calculator are best compared to 14.7% of choices made without the calculator. Results suggest that having a calculator and being paid depending on choices is additive, both improve the proportion of payoff maximizing choices in choice sets D to I, but are unsuccessful at improving choices in choice sets A, B and C. The proportion of payoff maximizing choices in these choice sets remains disappointing.

In **experiment 5** (*oath on truth*) subjects are asked to sign a truth-telling oath. All subjects except one signed the oath.[6] Table 3 presents the percentage of best choices in experiment 5. The results are unambiguous. The oath significantly increases the percentage of best choices compared to experiment 1: 78.3% compared to 56.3% ($p < .001$). The increase is due to improvements in choice sets A, B and C: the percentage of best choices increases by a factor 3, from 22.7% in experiment 1 to 76.5% in experiment 5 ($p < .001$). We observe no significant difference in choice sets D to I: 73% in experiment 1 and 78.8% when subjects are under oath ($p = .683$). Figure 1.e presents the percentage of best choices a subject makes, 40.9% of subjects make best choices in all 9 choice sets and 54.5% make at most one bad choice. The EDF in experiment 5 significantly first order dominates the EDF in experiment 1 ($p < .001$).

Response time in experiment 5 is significantly longer than experiment 1. The response time increase is comparable to the increase observed in experiment 3 – *paid*. Median total response time is 208 seconds in experiment 5 compared to 157 seconds in experiment 1 ($p = .008$) and 197 seconds in experiment 3 ($p = .400$). The EDF of response time in experiment 5 first order dominates that of experiment 1 ($p = .004$, see EDF of response time in experiments 1, 3 and 5 in appendix E).

The longer response time in experiment 5 has two sources. First, subjects making a best choice in choice sets A, B and C take longer in experiment 5 than in experiment 1: the median response time is 21 seconds in experiment 5 compared to 12 seconds in experiment 1. In experiments 1 and 3, best choices are made quicker than bad choices: median response time for best choices is 14 seconds compared to 21 seconds for bad choices. This is no longer the case in experiment 5: median response time for bad choices is now 19 seconds.

Second, response times are longer for both best choices and bad choices in choice sets D to I. Median response time is increased by 6 seconds to 18 seconds for bad choices (12 seconds in experiment 1 and 16 seconds in experiment 3) and by 4 seconds to 21 seconds for best choices (17 seconds in experiment 1 and 20 seconds in experiment 3).

The response time in experiment 5 shows that the truth-telling oath increases cognitive reasoning. One explanation is that subjects under oath dedicate more effort to making accurate choices. What remains striking is that monetary incentives in experiment 3 also increase cognitive reasoning, but without improving choices in choice sets A, B, and C. One explanation is that engaging subjects by an external non-monetary commitment device is more efficient than money (experiment 3) at fostering one's cognitive effort to perform the task accurately .

In **experiment 6** (*oath on task)* we explore whether the explanation that the oath improves choices because it fosters cognitive reasoning is reasonable. Subjects are asked to sign an *oath on task* and all subjects agree. The percentage of best choices presented in Table 3 shows that the task oath has only a small positive effect on choices compared to experiment 1: 63.7% compared to 56.26% ($p = .074$). This improvement comes from better choices in choice sets D to I (84.5% in experiment 6 vs. 73% experiment 1, $p = .071$). The task oath has no effect on choices in choice sets A, B and C (21.6% in experiment 6 vs. 22.7% in experiment 1, $p = .321$).

This result is confirmed at the subject level by the comparison of the EDF in experiments 1 and 6 (Figure 1.f): the EDF in experiment 6 does not first order dominate that in experiment 1 ($p = .536$).

Response times show that subjects take the task oath seriously. There is a significant increase in response time compared to experiment 1: median total response time is 237 seconds in experiment 6 compared to 157 seconds in experiment 1 ($p = .012$). The EDF of response time in experiment 6 first order dominates that of experiment 1 with $p < .001$. The response time increase is similar to that for the truth-telling oath (experiment 5): median

[6] Choices of the subject who did not signed are not dropped from the statistical analysis, i.e. we adopt an intention to treat strategy.

total response time is 208 seconds in experiment 5 (a response time of 237 seconds is not significantly different from 208, $p = .176$).

In experiment 6, best choices are made quicker than bad choices in choice sets A, B and C. Median response time of best choices is 13.5s, 18s and 5s, respectively. Median response time of bad choices in choice sets A, B and C is 19s, 37.5s and 26s respectively. This follows the pattern observed in experiments 1 (baseline) and 3 (paid), and is in contrast to experiment 5 (truth) in which best choices take longer. In choice sets D to I, individuals take more time to make best (25s) rather than bad (14.5s) choices – 17s and 12.5s in experiment 1 and 22s and 18s in experiment 5.

Response time in experiment 6 shows that the task oath fosters cognitive reasoning which improves choices made in choice sets D to I, but has no impact on choices in choice sets A, B and C. This suggests that it is not by fostering cognitive reasoning that the truth-telling oath improves decision making. The task oath may appear too abstract and singular, whereas the truth-telling oath is a real world institution with a moral content. The task oath may be too far from what one would encounter in the field and therefore it does not succeed in increasing the proportion of payoff maximizing choices.

In **experiment 7** (*oath on duty)* subjects are asked to sign an oath that invokes their sense of duty. All subjects except one signed the oath.[7] The oath of office has no overall effect on behavior. Table 3 shows that 61.6% of choices are best, which is not statistically different than experiment 1 ($p = .185$). As with the task oath, we observe, a small but significant increase in best choices in choice sets D-I (84.7% in experiment 7 vs. 73.0% in experiment 1, $p = .068$), but no effect in choice sets A, B and C (21.6% in experiment 7 vs. 22.7% in experiment 1, $p = .834$). At the subject level, the comparison of the EDF of the percentage of best choices in experiment 1 and 7 presented in Figure 1.f confirms that the oath of office does not improve choices in our setting.

Comparing the oath on duty to the oath on truth and the oath on task, we find that response time data shows that subjects take the oath on duty seriously. Median total response time in experiment 7 is 213s, significantly greater than in experiment 1 (157s) and similar to both other oath experiments. The EDF of total response time for each subject is presented in appendix G shows that the EDF in experiment 7 first order dominates the EDF in experiment 1 (KS bootstrap test, $p = .009$) and that it is similar to that of experiment 6. Response times for best choices are longer than for bad choices with the oath on duty, as with the oath on truth. Subjects making best choices take longer to choose (25s) than those who make bad choices (21s). Still, many subjects do not maximize their hypothetical monetary payoff in choice sets A, B and C.

In summary, our analysis of response time tells us that decisions in the most problematic choice sets (A, B and C) are not merely mistakes for two reasons. First, response time in problematic choice sets are longer for bad decisions. Second, devices that target cognitive effort show that people who spend more time deciding do not make better decisions in problematic choice sets. This suggests that people deliberately choose the wrong token after careful thinking, because in the truth-telling oath experiment, best choices take more time than bad choices in problematic choice sets (given that we also observed in experiment 3 a general increase in response time).

Conclusion

Discrete choice experiments (DCE) are a popular survey tool to elicit preferences for multi-attribute goods, e.g., improved environmental quality, food products created by new technologies. The many field applications maintain an implicit presumption that the people who answer DCE questions are telling the truth. But whether

[7] As in experiment 5, statistical analysis is carried out without dropping observations.

they actually are answering sincerely remains an empirical question – one that can be addressed in the laboratory in which researchers can control preferences through Smiths (1976) classic induced value design. Herein we do just that by examining truth-telling within a DCE framework under alternative institutional treatments. Our baseline results suggest truth-telling (or the lack of it) is a serious concern. Was weak truth-telling due to the complexity of the good? No, we find that helping respondents with a calculator had little or no effect on choices. Was weak truth-telling do to a lack of effort? No, we observe similar levels of weak truth-telling even after subjects signed an oath to "faithfully and conscientiously fulfill" his or her duties. Was weak truth-telling do to a lack of commitment to tell the truth? Yes, we find that the truth-telling oath was the only institutional rule that we considered that improved sincere behavior within the DCE setting.

These findings suggest truth-telling should be explicitly addressed and not implicitly assumed in field applications. Researchers could ask respondents to take a truth-telling oath prior to being interviewed. Alternatively, future research might consider whether weaker forms of commitment would be successful such as a preliminary pledge or even a signed agreement to tell the truth. One might counter our recommendation by arguing our induced value experimental design was too abstract. Basic math problems do not reflect real world goods and services. While we appreciate and understand this reasoning, we defend our approach based on two comments. First, this reasoning supposes that real choices exist and are observable. But in many situations, such as health care or environmental preservation, the DCE must be carried out before any policy exists, and no markets exist in which one could observe choices and competitive prices. Second, and most importantly, this reasoning presumes that real choices in the field fulfill the necessary rational requirements that underpin meaningful welfare estimates. As we have seen, in our basic task, monetary incentives alone do not guarantee that people choose their best option. Context matters. We believe our design is a necessary starting point to understand better whether preferences revealed in field DCE surveys are accurate enough for collective decision making.

References

[1] Carson, R. T. (2000) Contingent valuation: a user's guide. *Environmental Science & Technology* **34**, 1413–1418.

[2] Cummings, R. G & Taylor, L. O. (1999) Unbiased value estimates for environmental goods: A cheap talk design for the contingent valuation method. *American Economic Review* **89**, 649–665.

[3] Hainmueller, J, Hangartner, D, & Yamamoto, T. (2015) Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences* **112**, 2395–2400.

[4] Chetty, R. (2015) Behavioral economics and public policy: A pragmatic perspective. *American Economic Review* **105**, 1–33.

[5] Luchini, S & Watson, V. (2014) Are choice experiments reliable? evidence from the lab. *Economics Letters* **124**, 9–13.

[6] Krajbich, I, Bartling, B, Hare, T, & Fehr, E. (2015) Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nat Commun* **6**, –.

[7] Evans, A. M, Dillon, K. D, & Rand, D. G. (2015) Fast but not intuitive, slow but not reflective: Decision conflict drives reaction times in social dilemmas. *Journal of Experimental Psychology: General* **144**, 951–966.

[8] White, C. N, Curl, R. A, & Sloane, J. F. (2016) Using decision models to enhance investigations of individual differences in cognitive neuroscience. *Frontiers in Psychology* **7**.

[9] Greiner, B. (2015) Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association* **1**, 114–125.

[10] Fischbacher, U. (2007) z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* **10**, 171–178.

[11] Taylor, L. O, McKee, M, Laury, S. K, & Cummings, R. G. (2001) Induced-value tests of the referendum voting mechanism. *Economics Letters* **71**, 61–65.

[12] Mazzotta, M. J & Opaluch, J. J. (1995) Decision making when choices are complex: A test of heiner's hypothesis. *Land Economics* **71**, 500–515.

[13] Swait, J & Adamowicz, W. (2001) Choice environment, market complexity, and consumer behavior: A theoretical and empirical approach for incorporating decision complexity into models of consumer choice. *Organizational Behavior and Human Decision Processes* **86**, 141–167.

[14] DeShazo, J & Fermo, G. (2002) Designing choice sets for stated preference methods: The effects of complexity on choice consistency. *Journal of Environmental Economics and Management* **44**, 123–143.

[15] Olof, J.-S & Henrik, S. (2008) Measuring hypothetical bias in choice experiments: The importance of cognitive consistency. *The B.E. Journal of Economic Analysis & Policy* **8**.

[16] Camerer, C. F & Hogarth, R. M. (1999) The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty* **19**, 7–42.

[17] Jacquemet, N, Joule, R.-V, Luchini, S, & Shogren, J. F. (2013) Preference elicitation under oath. *Journal of Environmental Economics and Management* **65**, 110–132.

[18] de Magistris, T & Pascucci, S. (2014) Does "solemn oath" mitigate the hypothetical bias in choice experiment? a pilot study. *Economics Letters* **123**, 252–255.

[19] Jacquemet, N, James, A, Luchini, S, & Shogren, J. (2011) Social psychology and environmental economics: a new look at ex ante corrections of biased preference evaluation. *Environmental & Resource Economics* **48**, 411–433.

[20] Mazar, N, Amir, O, & Ariely, D. (2008) The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research* **45**, 633–644.

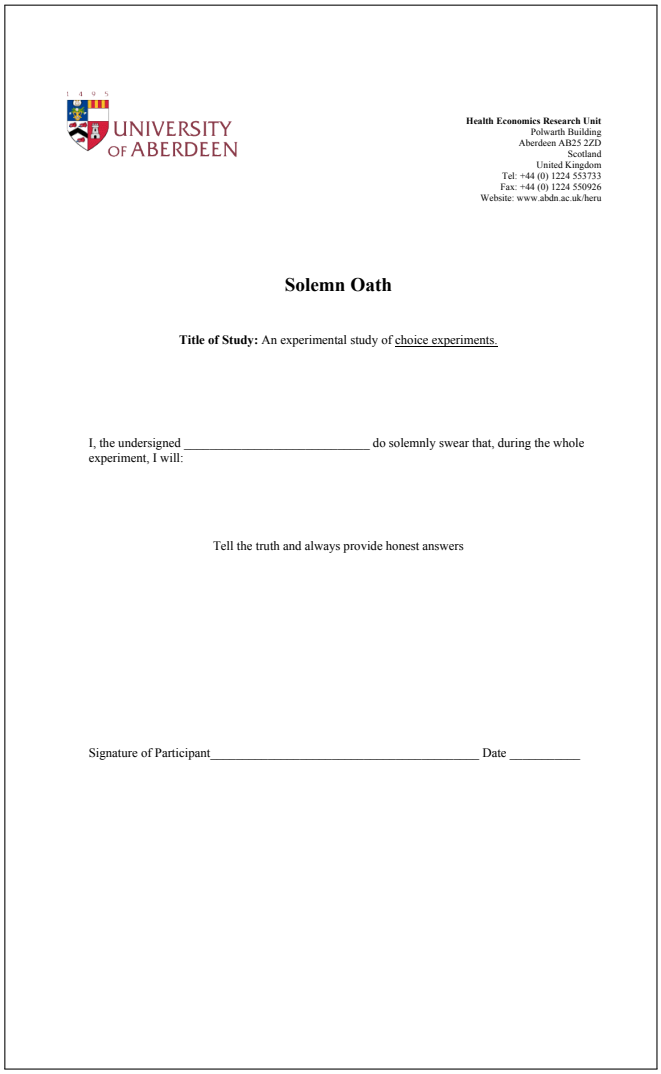
[21] Rubinstein, A. (2007) Instinctive and cognitive reasoning: A study of response times. *Economic Journal* **117**, 1243–1259.

[22] Abadie, A. (2002) Bootstrap tests for distributional treatment effects in instrumental variable model. *Journal of the American Statistical Association* **97**, 284–292.

[23] Sekhon, J. (2011) Multivariate and propensity score matching software with automated balance optimization. *Journal of Statistical Software* **42**, 1–52.

Appendix

A Oath form used in experiment 5



UNIVERSITY OF ABERDEEN

Health Economics Research Unit
Polwarth Building
Aberdeen AB25 2ZD
Scotland
United Kingdom
Tel: +44 (0) 1224 553733
Fax: +44 (0) 1224 550926
Website: www.abdn.ac.uk/heru

Solemn Oath

Title of Study: An experimental study of choice experiments.

I, the undersigned _____ do solemnly swear that, during the whole experiment, I will:

Tell the truth and always provide honest answers

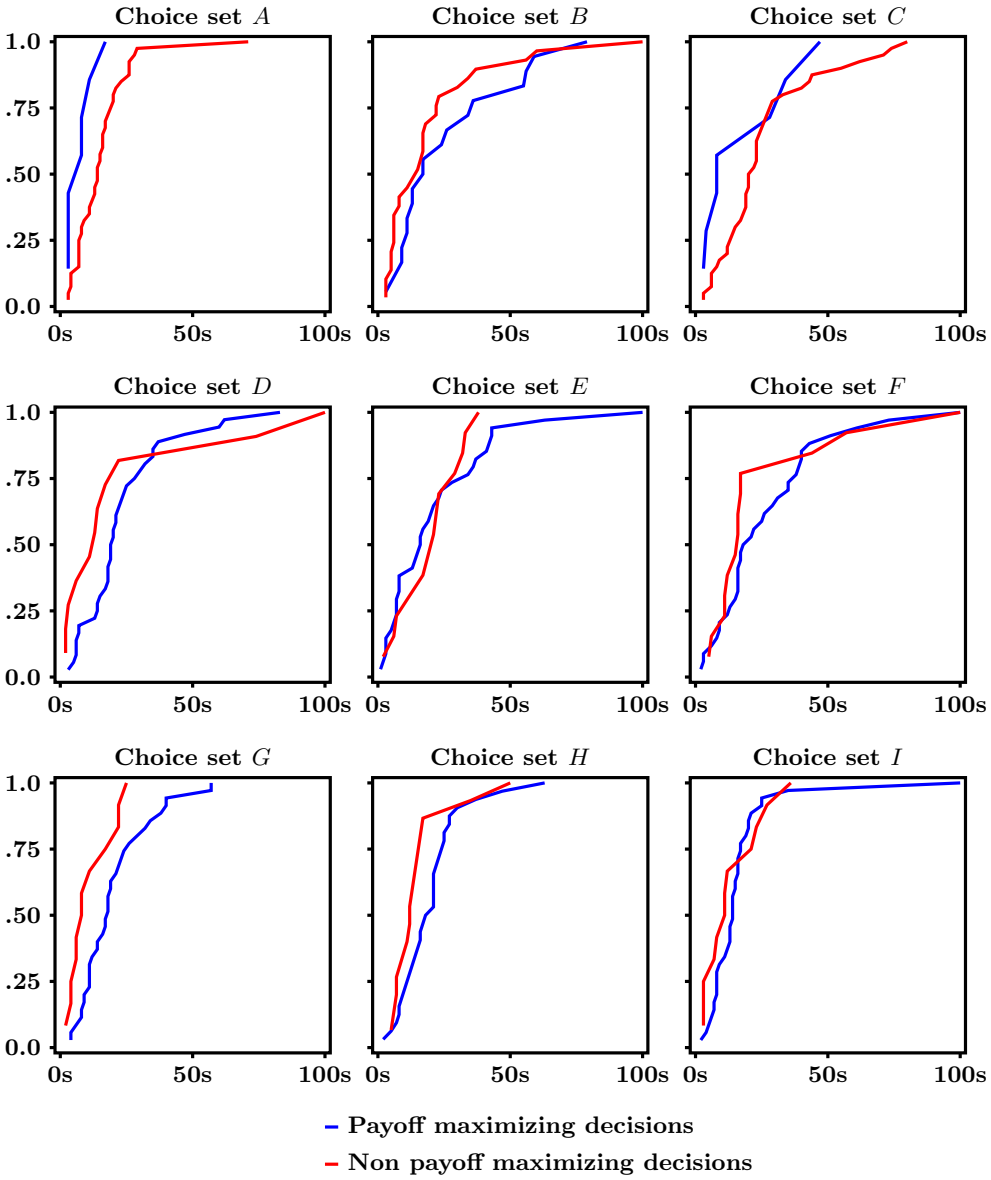
Signature of Participant_____ Date _____

B Payoff maximizing decisions by round and experiments

Round	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Baseline	42.5%	51.1%	57.4%	70.2%	61.7%	46.8%	55.3%	61.7%	61.7%
Calculator	60.9%	65.2%	58.7%	67.4%	60.8%	56.5%	56.5%	67.4%	60.8%
Paid	53.7%	55.5%	68.5%	50.0%	61.1%	64.8%	59.3%	70.4%	55.5%
Paid & Calculator	74.4%	64.1%	56.4%	66.7%	69.2%	56.4%	69.2%	61.5%	66.7%
Oath	72.7%	77.3%	84.0%	75.0%	77.3%	79.5%	77.3%	77.3%	81.8%

C Response time by choice set in experiment 1

Empirical distribution functions of response time in experiment 1 are plotted by choice set in the figure below. The x-axis is labeled in seconds and ranges from 0 second to 100 seconds. For the sake of visualization, 6 observations –1.4% of the sample– greater than 100 seconds were dropped. The red line corresponds to decision time of non payoff maximizing decisions and the blue line to payoff maximizing decisions. In choice sets A and C, the red line appears to the right of the blue line: EDF of response time of non payoff maximizing decisions first order dominate (FOD) the EDF of response time of payoff maximizing decisions. That is, a non payoff maximizing decision take more time than a payoff maximizing decision. For all other choice sets, response time EDF are either similar or the EDF of non payoff maximizing decisions appear to the left of that of payoff maximizing decisions.

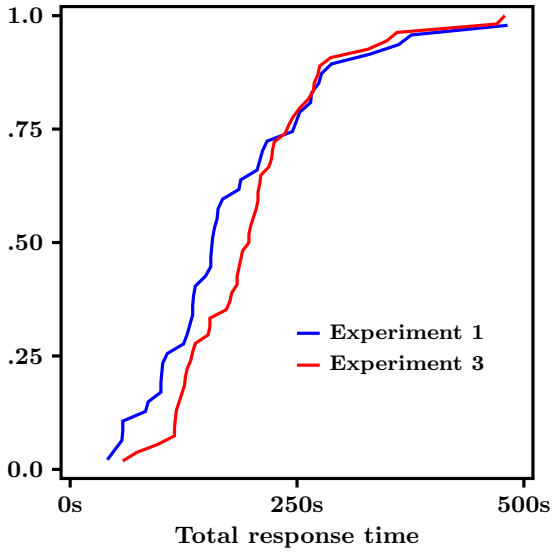


For each choice set, we perform a Kolmogorov-Smirnov bootstrap distribution test to assess whether first order stochastic dominance (in one way or the other) is statistically significant or not. KS bootstrap tests indicate that FOD of response time of non payoff maximizing decisions over response time of payoff maximizing decisions is statistically significant for choice set A and C with $p = .032$ and $p = .069$ respectively. FOD of response time of payoff maximizing decisions over non

payoff maximizing decisions is statistically significant for choice sets B, D, G and H with $p = .075$, $p = .041$, $p = .013$ and $p = .021$ respectively. The null of non FOD cannot be rejected for choice sets E ($p = .391$), F ($p = .114$) and I ($p = .127$).

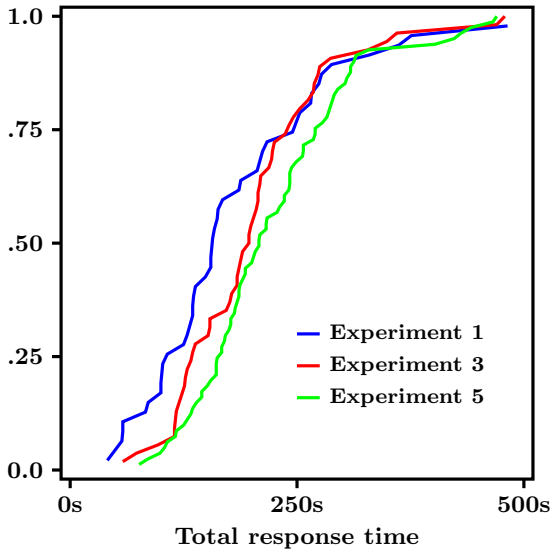
D Response time in experiment 1 and experiment 3

In the Figure below, we compute for each subject the time taken to answer all 9 choice sets (total response time) and plotted the EDF of total response time in experiment 1 and experiment 3. Note that we dropped one subject in experiment 1 with a response time of 765 seconds to make the Figure easier to read. The KS bootstrap test presented in the text is carried out without dropping this subject.



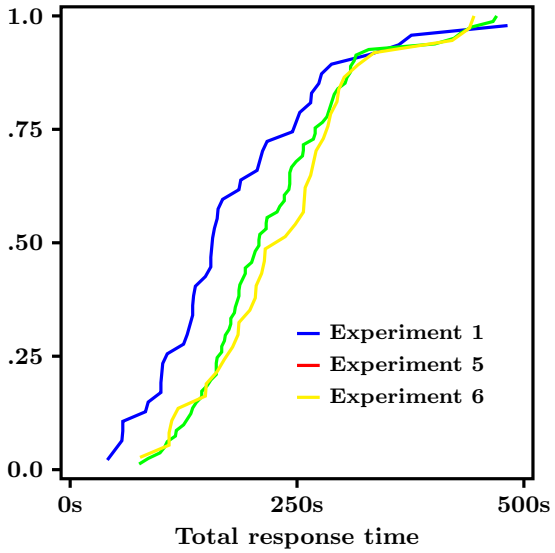
E Response time in experiments 1, 3 and 5

In the Figure below, we compute for each subject the time taken to answer all 9 choice sets (total response time) and plotted the EDF of total response time in experiments 1, 3 and 5.



F Response time in experiments 1, 5 and 6

In the Figure below, we compute for each subject the time taken to answer all 9 choice sets (total response time) and plotted the EDF of total response time in experiments 1, 5 and 6.



G Response time in experiments 1, 5 and 7

In the Figure below, we compute for each subject the time taken to answer all 9 choice sets (total response time) and plotted the EDF of total response time in experiments 1, 5 and 7.