Santiago Burone, Lukas Leitner

Correcting for Starting Point Bias in the Elicitation of Willingness to Pay for Health

Working Paper

No. 25/01

February 2025

University of Antwerp
Herman Deleeck Centre for Social Policy
https://www.uantwerpen.be/en/research-groups/csb/

Correcting for Starting Point Bias in the Elicitation of Willingness to Pay for Health

Santiago Burone[1]   Lukas Leitner[2]

[1,2]Herman Deleeck Centre for Social Policy, University of Antwerp

[1]Paris School of Economics

Working Paper No. 25/01

February 2025

Abstract

Willingness to pay (WTP) has become an important tool in economic analysis, despite the difficulty to obtain reliable estimates. This paper investigates the occurrence of starting point bias when eliciting WTP for health, a domain where this phenomenon has received limited attention, and illustrates its effect on equivalent consumption, a preference-based well-being measure. In an online experiment, three experimental groups responded to two dichotomous choice questions, with varying initial bids. The treatment groups then provided exact estimates for their WTP in an open-ended question. We find strong evidence for the existence of the bias using both non-parametric and parametric tests, and estimate a sizeable overall effect. Different parametric specifications yield point estimates between 29 and 43 percent for the first bid, whereas the effect of the second bid, which we estimate using an instrumental variable approach, is not statistically different from zero. We propose two *ex post* approaches to address this effect when using WTP data for interpersonal well-being comparisons. Although the percentage of rankings reversals is relatively small across all feasible comparisons, it becomes notable when examining comparisons for individuals within the same consumption deciles.

Keywords: Starting point bias; Willingness to pay; Dichotomous choice; Equivalent consumption.

Acknowledgements: The authors would like to thank FWO/FNRS for their support through the EOS project 30544469. This paper has benefited substantially from comments made by Koen Decancq and participants at seminars in Antwerp and Paris.

Corresponding author: Lukas Leitner (lukas.leitner@uantwerpen.be), Sint-Jacobstraat 2, 2000 Antwerp, Belgium.

1 Introduction

The literature on the measurement of well-being has reached a wide consensus over the past decades that well-being should be viewed as a multidimensional phenomenon (see, for instance, the influential report by Stiglitz, Sen, and Fitoussi, 2009). The set of well-being dimensions that should be included and the method of aggregating these dimensions into one measure, however, depends on a number of debatable normative considerations and may be context-specific. Yet, there is also a wide consensus that health is one of the most important determinants of well-being, be it self-reported or constructed by expert opinions. In this paper, we follow the stream of literature that aggregates different life dimensions, in our case consumption and health, on the basis of individual preferences. The equivalent consumption measure achieves this by subtracting the willingness to pay (WTP) for perfect saturation of the health dimension from the actual level of consumption. The resulting outcome bundle yields the same utility as the actual outcome bundle and can only vary in the level of consumption; thus, equivalent consumption is a cardinal measure of well-being that is interpersonally comparable.[1] A number of approaches can be applied to elicit individual WTP for the equivalent consumption measure (see Da Costa et al., 2024, for an overview). Here, we take a closer look at the contingent valuation (CV) approach, where people are asked directly for their WTP for a clearly defined hypothetical change of a given good or circumstance. Our aim is to understand the robustness of the CV approach in estimating equivalent consumption using health as a life dimension.

In their seminal paper, Tversky and Kahneman (1974) describe several cognitive biases which occur when people imagine, memorise, or estimate things. The authors emphasise that "[t]hese biases are not attributable to motivational effects such as wishful thinking or the distortion of judgements by payoffs and penalties" (p. 1130), implying that researchers cannot trust blindly in statements made in surveys, even when the respondents have the best intentions to answer correctly. One of these biases is almost omnipresent and persists even when using careful elicitation techniques: the anchoring effect. It refers to the human tendency to anchor numeric valuations to salient reference points, so-called "anchors", even when these are clearly unrelated to the question at hand. Ariely, Loewenstein, and Prelec (2003) demonstrate this phenomenon in a well-known experiment by first asking respondents whether they would buy an ordinary consumer product if its price was equal to the last two digits of their social security number in US$, followed by asking for their exact WTP.

This plays an important role for the validity of preference estimates given by the CV approach. For applications on abstract and intangible goods, such as cultural assets, environment protection, or national defence, the NOAA Panel (Arrow et al., 1993) recommends supporting respondents in their decision-making process by asking referendum-like dichotomous choice (DC) questions.[2] Since DC questions necessarily require a first bid (or "starting point"), this approach is prone to the so-called "starting point bias" (SPB), a variant of the anchoring effect.

This effect has been studied extensively in the environmental economics literature. The first suggestions of a hypothetical SPB in CV studies date back to the 1970s. In efforts to elicit the willingness to pay (WTP) for environmental goods, DC approaches (see, for instance, Randall, Ives, and Eastman, 1974; Brookshire, Ives, and Schulze, 1976; Randall, Grunewald, et al., 1978)

[1] See Decancq, Fleurbaey, and Schokkaert (2015) for a theoretical foundation of the measure and Decancq and Nys (2021) for an application using an iterative DC approach in the income-health space.

[2] Polome, van der Veen, and Geurts (2006) examines the impact of including the term "referendum" in CV-DC questions, as suggested by the NOAA panel, on survey results. The study focuses on the Netherlands and reveals notable differences in outcomes based on the use of the term. The NOAA panel advocates for the referendum format, arguing that respondents are generally familiar with it. While this might be true for the case of the US, the authors caution that in contexts where referendums are less common, including the term might lead respondents to perceive the issue as more significant than intended.

yielded seemingly different results than other CV approaches (such as Hammack and Brown Jr., 1974). This discrepancy instigated the first systematic studies to detect the SPB in the 1980s (see Rowe, d'Arge, and Brookshire, 1980; Thayer, 1981; Boyle, Bishop, and Welsh, 1985; Samples, 1985), and the first studies to model and quantify the bias in the 1990s (e.g. Boyle, 1990; Cameron and Quiggin, 1994; Li and Mattsson, 1995; Hanemann and Kanninen, 1996; Herriges and Shogren, 1996; Alberini, Kanninen, and Carson, 1997). Finally, with the emergence of more powerful computers, semi-non-parametric models (Cooper, 2002) and Bayesian methods (R. León and C. J. León, 2003; Araña and C. J. León, 2007) were first tested in the 2000s.

Despite the large scope of the existing literature on SPB, some questions have not been answered satisfactorily. First, there is only limited research on the presence of SPB in other fields of economics. Onwujekwe and Nwagbo (2002) and Heinzen and Bridges (2008) investigate its effects on CV approaches to estimate the value of health treatments. While Onwujekwe and Nwagbo find no evidence for SPB in a CV survey on insecticide-treated nets, Heinzen and Bridges find a significant and sizeable effect in a survey on the economic value of pneumococcal vaccines. However, it is not evident how informative these findings are for SPB in DC studies on environmental goods, since the good health is of a different nature than environmental protection. Health is a private good that matters to most (if not all) individuals, regardless of how altruistic or well-informed someone is. Consequently, one could argue that people have more clearly defined preferences for their health than for environmental goods, such that their sensitivity to SPB could be lower in this context. Thus, the question at hand is whether the SPB is also present in the elicitation of WTP for health, and how large it is compared to the more abstract good of environmental protection.

Second, there is no conclusive evidence about heterogeneity in the sensitivity to the bias. Aprahamian, Chanel, and Luchini (2007) investigate this question by modelling SPB as a random value which is drawn from a distribution that may vary between socio-demographic groups. In their application on WTP for less pollution, they find evidence for heterogeneity in the bias; however, it is not apparent that the findings are robust when the estimators' imprecision is taken into account.[3] Our ambition in this paper is to identify heterogeneity in the bias sensitivity by simpler, more robust means, at the potential cost of overlooking patterns in higher moments.

Third, even though a large number of contributions aim at detecting and quantifying SPB, we are aware of only one study that offers a model to correct for the bias. Liou (2015) uses the baseline model proposed in Herriges and Shogren (1996) and inverts the equation which relates true and stated WTP, while simultaneously using a Tobit model to account for censored protest votes. This way, true WTP can be expressed as a function of stated WTP and the estimated coefficient for SPB. However, we see two issues that are not discussed in the study: first, the effect of the second bid on stated WTP is implicitly assumed to be zero, and second, the estimated true WTP may be biased by the elicitation procedure itself, and can as such not be universally true. Particularly the second issue poses a problem for the external validity of bias-corrected CV surveys; if respondents generally respond differently to open-ended (OE) questions after seeing two DC questions—for instance, due to learning effects or yea-saying bias—we have to

[3] In their pioneering study, the authors use a beta distribution to model the random bias size. The two shape parameters of the beta distribution are estimated using interaction terms with two socio-demographic characteristics (gender and income), allowing for heterogeneity in the bias sensitivity. Since the mean and the variance of the beta distributions each depend on both shape parameters, it is not evident how they are affected by the imprecision in the parameter estimates. The estimates appear in the denominator of both the mean and the variance, and may additionally be correlated with each other. Even less intricate cases can be challenging for statistical inference; see, for instance, Hole (2007), who compares techniques to find confidence intervals for (one-dimensional) normal ratio distributions.

reconsider which frame[4] to use as a reference for the unbiased, true WTP, and whether such a true WTP exists at all.

To investigate these questions, we conducted an experiment in the form of an online survey. The respondents first answered two modules containing questions on health (physical and mental) and consumption, such that we can construct a health index and calculate monthly consumption. Afterwards, the respondents saw two DC questions on whether they preferred their own life, with their current levels of health and consumption, or an alternative life, with perfect health and a lower level of consumption, followed by an OE question asking for the minimum value of consumption at which they would prefer the alternative life. By randomly assigning the respondents to a low, medium, or high starting point, we created three treatment groups.

We present three complementary methods to identify the SPB: a non-parametric method, a parametric model of the treatments, and a parametric model of the bids size. For the model on the bid sizes, we employ an instrumental variable (IV) strategy to distinguish the effect of the second bid from the effect of first. Since the size of the second bid depends both on the size of the first bid and on the first choice made by the respondent, which in turn depends on the respondent's true WTP, it is an endogenous variable.[5] We address this endogeneity issue by using a novel set of instruments related to the visible exposure of the bids to the respondents.

Our results present robust evidence for SPB in the estimation of WTP for health. The evidence suggests that DC responses are influenced by a complex interplay of mechanisms, therefore solely analysing DC responses could yield an incomplete understanding of its magnitude. By further analysing OE responses, we can detect the influence of the initial bids more accurately, showing that they act as salient anchors in the distribution of responses. Our parametric analysis confirms this. Even in a simple model with few parametric assumptions, we observe that most socio-demographic groups react similarly—or at least not significantly differently—to varying the bids presented in the DC questions. This finding is encouraging, as it suggests not many groups are disproportionately affected by the SPB, which could otherwise compromise the validity of WTP estimates. However, we find evidence that highly educated individuals and those who declare being able to imagine the hypothetical scenarios posed in the DC questions are less influenced by high initial bids.

Efforts to assess SPB by regressions on WTP comprise the challenge of subsequent bids being endogenous to the dependent variable. This has led previous studies to overlook the second bid's effect during estimation. Leveraging our dataset, we use an instrumental variable strategy to include both bids in the model, providing innovative evidence that when accounting for endogeneity the second bid's influence is smaller than the first bid's and not statistically different from zero. We estimate the first bid's effect to range from 0.42 (with a standard error of 0.10) when including the instrumented second bid in the model, and to 0.30 (with a standard error of 0.05) when omitting the second bid.

Finally, we propose an approach to account and correct for the bias in the construction of the equivalent consumption measure. We compute both observed equivalent consumption by subtracting stated WTP from consumption, and hypothetical equivalent consumption under each treatment; this hypothetical equivalent consumption uses the estimated WTP an individual would have stated had she been in a different treatment group and seen different bids. We then construct well-being rankings based on observed and hypothetical equivalent consumption to illustrate the

[4] The term "frame" denotes the set of "observable information that is irrelevant in the rational assessment of the alternatives, but nonetheless affects choice". This definition is taken from the abstract of Salant and Rubinstein (2008). Another study, which was published shortly after (Bernheim and Rangel, 2009), uses the term "ancillary condition" in a similar framework of generalised choice situations.

[5] Note that this is a problem inherent to the use of follow-up DC questions, and has, to the best of our knowledge, not been discussed before.

impact of disregarding SPB. The share of rankings which is reversed when using hypothetical WTP depends on the size of the SPB, but also on whether stated WTP represents a relevant share of consumption; if stated WTP is small compared to consumption, equivalent consumption is almost unaffected by the bias, and equivalent consumption rankings are similar to rankings based on consumption alone. Further, the ranking of hypothetical equivalent consumption may be incomplete; that is, if one individual has a higher hypothetical equivalent consumption than another for some, but not for all treatments, the two individuals cannot be ranked, unless an additional ranking criterion is chosen. We inspect the level of incompleteness for the super safety principle proposed by Fleurbaey and Schokkaert (2013), and for a treatment-wise comparison.

For the calculation of hypothetical treatments, we propose two *ex post* approaches, a non-parametric and a parametric one. The outcome from this exercise suggests that, while the overall effect of disregarding SPB on ranking reversals is small, the ordering changes significantly when focusing on comparisons of individuals with similar levels of consumption. The extent of this re-ranking is more pronounced for the parametric approach, despite the relatively small degree of heterogeneity in bias sensitivity.

The paper is organised as follows: Section 2 describes the experimental design and the data; Section 3 describes our non-parametric analysis; Section 4 progresses to a parametric analysis; Section 5 explores the implications of SPB in utilising WTP for well-being comparisons; and Section 6 concludes.

2 Online Survey Experiment

In March 2023, we conducted an online survey using the platform Qualtrics. The survey featured an experimental treatment and was administered to a sample of 1494 Spanish respondents, who received a reward according to Qualtrics' internal incentive scheme.[6] To test for heterogeneity in the answering behaviour, quotas were applied on the respondents' age and level of education.[7] Further, respondents were required to have a stable source of income, such that a minimum level of consumption and financial literacy could be presumed. While the sample is not representative of the Spanish population, it is balanced across treatments (see Table 10).

The survey started with a module on self-assessed health (mental and physical), followed by a module asking for the respondents' consumption expenses for several items.[8] The health index was summarised in an index from 0 (worst health) to 100 (perfect health), and respondents were asked to confirm or adjust the health index and consumption level estimate.[9] In the experimental module, the respondents stated whether they would prefer to live their current lives, with their current health circumstances and their current level of consumption, or an alternative life, in which all their health problems have disappeared, but their personal consumption level has been reduced to a given share of their consumption level. This DC question was then repeated once with an adjusted value of consumption in the alternative life, depending on the previous answer. Afterwards, respondents were asked to state the minimum consumption level at which they

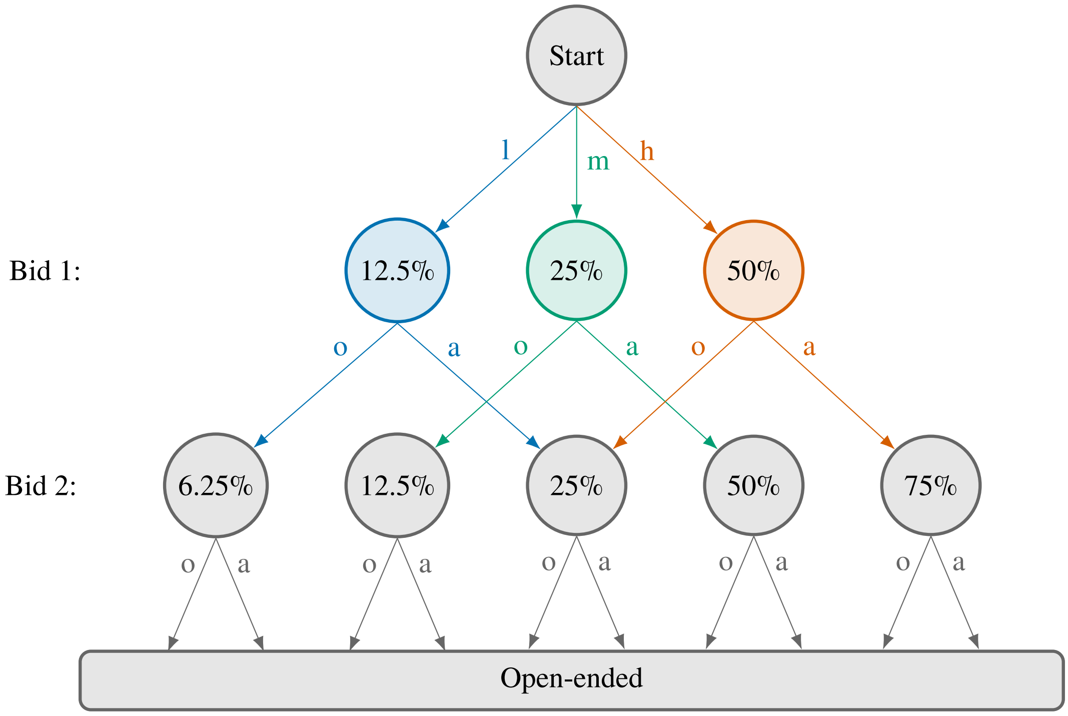
[6] The questionnaire has received ethical approval from the Ethics Committee of Social Sciences and Humanities at the University of Antwerp with the file number SHW_2022_171_1). The questionnaire can be found online.

[7] Respondents were required to be between 25 and 64 years old, with equal proportions in the groups of 25–34 years, 35–44 years, 45–54 years, and 55–64 years. Additionally, we required that the proportions for none or primary education, secondary education, and tertiary education were equal. This quota was later lifted, due to difficulties to reach the requested number of respondents without secondary or tertiary education.

[8] The questions included in the health module were adapted from the 12-Item Short Form Health Survey (SF-12) with modifications.

[9] If the respondents indicated that they were not living alone, we further asked how the usage of shared items was allocated within their household, such that we could estimate a value for their own monthly consumption.

Figure 1 Diagram of treatments in starting point bias experiment



would be willing to accept the alternative life, thus stating an exact willingness to pay for perfect health in an OE format question.[10]

The only difference between our experimental treatments is the amount by which their consumption is reduced in the alternative lives. In the first DC question, this value could either be 12.5%, 25%, or 50% lower than their actual level of consumption.[11] We will refer to these treatments as the "low", "medium", and "high" treatment. Depending on whether the respondents indicated to prefer their own life "o" or the alternative life "a", the bid was adjusted for the second DC question.[12] A scheme with the possible bids for each treatment and response can be found in Figure 1. After answering the DC questions, every respondent saw the same OE question, in which they provided an exact value of WTP for perfect health. If the stated WTP from the OE question was negative or outside the range given by the DC questions, respondents were asked to explain or correct their answers.

After applying some essential quality filters on the full sample with 1494 observations, a sample with 1184 observations remains. Figure 7 presents a simplified illustration of the

[10] The DC questions and the OE question can be found in Appendix A.1.

[11] The values were calibrated with a pilot sample of 500 respondents who only saw the OE question, and roughly represent the 20th, 50th, and 80th percentile of the WTP distribution. These percentages emerge from a trade-off between statistical power and credibility; on one hand, they create sufficient variation in the starting points to reach a high level of precision in the estimates, and on the other hand, they may still seem plausible to most respondents. The rounding to 1/8, 1/4, and 1/2 further connects to Decancq and Nys (2021), who use an adaptive DC mechanism that iteratively halves the remaining interval. The pilot sample was not balanced with the treatment groups in several socio-economic characteristics, even though the pilot study was conducted only one month earlier, and can hence not be used as another treatment here.

[12] If respondents from the medium treatment group answered "I don't know" to the first DC question, a tie-breaker was applied to determine if the second bid would be higher or lower. For respondents from the other two treatment groups, the second bid was set to 25%, i.e. the first bid of the medium treatment group.

Table 1 Consistency between DC and OE responses by treatment group

Treatment	Low	Medium	High
DC inconsistent	6	10	6
	1.50%	2.56%	1.52%
Refusal WTP	12	8	13
	3.00%	2.05%	3.30%
WTP < 0	60	38	52
	15.00%	9.74%	13.20%
Outside DC bounds	111	135	127
	27.75%	34.62%	32.23%
Consistent	211	199	196
	52.75%	51.03%	49.75%
Total	400	390	394

filtering process.[13] To conduct the analysis of the DC questions, an additional 22 observations with contradictory responses are discarded.[14] Table 1 shows the frequencies of consistent or inconsistent responses by treatment group. We find no indication of a pattern in inconsistent responses to the DC questions across treatments. For the analysis of the OE responses, only observations with WTP ≥ 0 are considered, meaning that we discard 186 observations with negative WTP or refusals to state any WTP.[15] Respondents could correct their responses if the DC responses contradicted each other, if the value from the OE response lay outside the upper and lower bounds given by the DC responses, or if it was negative. For 51.2% of the respondents in the filtered sample responses were consistent, and therefore they were not presented with the correction module.[16]

In addition to the health, consumption, and experimental module, we asked for socio-demographic characteristics (household size, gender, education, nationality, civil status, and

[13] Responses with a total answering time of less than four minutes were discarded automatically, resulting in an average answering time of about 15 minutes. The time used on specific screens, as well as the time to the first and the last mouse click, were also utilised to filter out low-quality responses. Observations in which the respondents spent less than 6 seconds on the first DC question, 3 seconds on the second DC question, and 6 seconds on the OE question (*ergo* 15 seconds for the experimental module in total) or in which the respondents stated to be in perfect health or to have a consumption of less than 200€ per month were discarded. The filters on time use in the experimental module are calibrated using two variables from the health module, which are strongly negatively correlated ("feeling weary and blue" versus "feeling energetic"). For respondents who stay below the minimum time, this correlation is close to 0, indicating that these respondents answered at random.

[14] Note that the DC responses can only contradict when the respondent either responded "I don't know" to the first DC question and then chose the alternative (own) life at a higher (lower) second bid, or made a mistake when correcting the responses to the DC questions.

[15] The frequency of negative WTP appears to be substantially lower for respondents in the medium treatment group; however, there are no statistically significant differences (a chi-square test comparing the frequencies of negative with positive WTP across treatments yields a p-value of 0.0782). Similarly, it appears that respondents in the medium treatment group tend to give OE responses that lie outside the DC bounds more frequently, but the difference is not statistically significant either (with a p-value of 0.2513 for the chi-square test).

[16] We consider these corrections to be driven by mistakes due to inattention and therefore use the corrected response in our analysis. However, one could be concerned that receiving such information might introduce an additional type of bias. We performed a sensitivity analysis using the original OE values, finding consistent results. If corrections were introduced, we asked respondents the reason for it; most indicated they had made a mistake. However, 101 respondents stated they had changed their minds between responses. We performed a sensitivity analysis keeping the original responses for those cases, finding again qualitatively robust results. For the parametric analysis, these additional robustness checks are provided in the Appendix (Section A.6).

employment), the locus of control, evaluation of the survey, financial awareness, satisfaction with life in general, health, and consumption, and how certain the respondents were with their responses. In the parametric analysis, we explore whether any of these variables may help in explaining which respondents are more vulnerable to SPB.

3 Non-Parametric Analysis

In the first part of the analysis, we use non-parametric methods to test whether the treatments had an effect on stated preferences. The null hypothesis is the following: if the respondents answered truthfully and without any behavioural bias, the treatments did not have an impact on the preferences elicited by the DC and OE questions. If there are (significant) differences, this indicates that the starting point did have an effect; however, to determine whether this effect can be attributed to the SPB (in contrast to, for instance, acquiescence bias), we must exclude the possibility that respondents were careless and simply chose their own or the alternative life with certain probabilities, regardless of the bid. We use “WTP” to refer to the absolute value of consumption that individuals are willing to forgo in exchange for perfect health, and “wtp” to refer to WTP as a proportion of consumption (where $0 \leq \text{wtp} \leq 1$). First, we examine the DC questions.

3.1 Dichotomous choice questions

Figure 2 displays the responses to the DC questions by treatment group. The histograms are constructed as follows: using the DC responses, we determine the lower and upper bound for the wtp of each respondent. If a respondent stated “I don’t know”, the bounds are unaffected. In case the respondent never chose the alternative life, we use 0 as the lower bound for her wtp, and in case she never chose her own life life, we use 1 as the upper bound. To obtain the density, the probability mass of each observation is then uniformly distributed over the interval between the lower and upper bounds.[17]

It is apparent that the densities in Figure 2 are not equal across treatments, and that respondents from the high treatment group are more likely to accept higher bids than respondents from the other treatment groups. Further, for respondents from both the medium and the high treatments, the density in the interval $[0.25, 0.5]$—where we expected to find the mean and median wtp—is lower than the density anywhere else. Yet, the figure has to be interpreted with care, since the bins are not identical across treatments.

We present three simple non-parametric test that make use of the fact that respondents from different treatment groups may have seen the same bid(s) in a different order. First, all treatment groups are compared at the initial bid of the medium treatment group ($b = 0.25$). Respondents from the low treatment group saw this as the second bid if they had chosen the alternative life in the first DC question (see Figure 1). If a respondent from the low treatment group chose her own life in the first DC question, we can implicitly assume (by monotonicity) that she would also have chosen her own life at any higher bid in the second DC question. Vice versa, the respondents from the high treatment group were also able to state a wtp higher or lower than 0.25, be it implicitly or explicitly. If the null hypothesis holds, the probability of choosing their own or the alternative life for a given bid is equal for all respondents. Thus, we check whether the choices at $b = 0.25$ vary significantly across treatment groups.

[17] Since observations with “I don’t know” responses span more than one bin, this procedure does not yield a “clean” histogram; however, with only around 9% of the respondents saying “I don’t know” at least once, the distortion is marginal. To improve readability, the vertical axis is limited to a value of 3; this cuts off the first bin of the low treatment group, which reaches a density of around 4.6.

Figure 2 Histograms of wtp elicited with DC questions

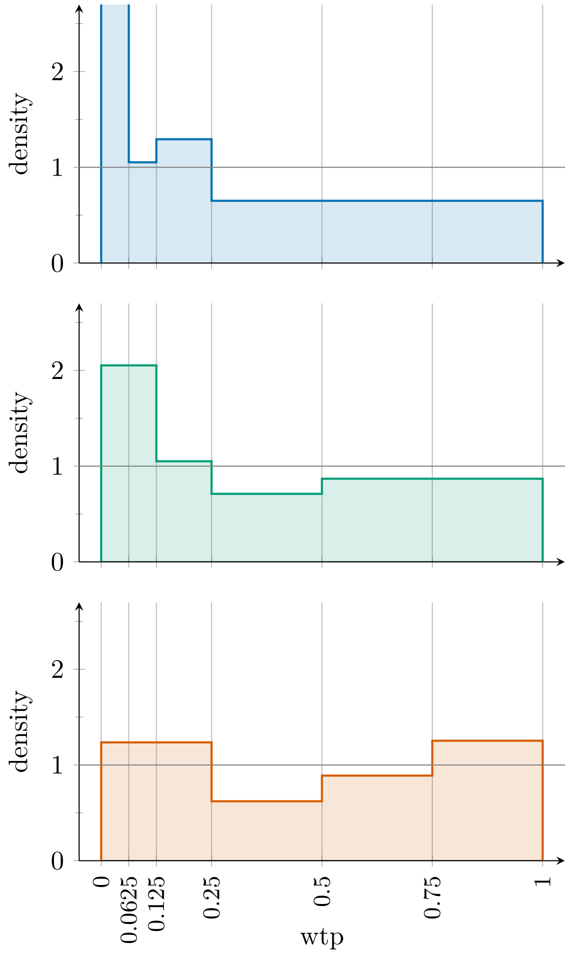


Table 2 Frequencies of wtp > 0.25 by treatment group

Treatment	Low	Medium	High
wtp < 0.25	196 53.55%	135 39.82%	108 30.68%
wtp > 0.25	170 46.45%	204 60.18%	244 69.32%
Total	366	339	352

Table 3 Frequencies of wtp in low and medium treatment group

Treatment	Low	Medium
wtp < 0.125	130	92
	36.72%	27.22%
wtp ∈ [0.125, 0.25]	54	42
	15.25%	12.43%
wtp > 0.25	170	204
	48.02%	60.36%
Total	354	338

Table 4 Frequencies of wtp in medium and high treatment group

Treatment	Medium	High
wtp < 0.25	135	108
	41.93%	32.14%
wtp ∈ [0.25, 0.5]	47	43
	14.60%	12.80%
wtp > 0.5	140	185
	43.48%	55.06%
Total	322	336

Table 2 shows that respondents from the high treatment group choose the alternative life more often than respondents from the other two treatment groups, and respondents from the medium treatment group choose the alternative life more often than respondents from the low treatment group. These differences are large in size, and a joint chi-squared test confirms that the overall correlation is highly significant (p-value < 0.000).

When comparing the low and the medium treatment group, we can perform the same test for two bids at the same time. In both groups, respondents were able to indicate a wtp below 0.125, between 0.125 and 0.25, or larger than 0.25. Similarly, the medium and the high treatment group can be compared using the frequencies of respondents indicating a wtp below 0.25, between 0.25 and 0.5, or larger than 0.5. Again, if the null hypothesis holds, we expect no significant differences in the response behaviour between the treatment groups.

Tables 3 and 4 show the results of these comparisons.[18] In both cases, the higher starting point is correlated with a higher stated wtp (these correlations are significant at the 2.5%-level).[19] Interestingly, it appears that the frequencies of choosing the alternative life is only weakly affected by the bids; when moving from Table 3 to Table 4 the frequencies in each line and row do not change much, even though the bids double.

This raises the question whether the correlation between the first bid and stated wtp is mainly driven by respondents choosing one option or the other with a certain probability, regardless of

[18] Note that the frequencies in both tables do not need to add up to the same total number, since "I don't know" answers may be excluded. For instance, if a respondent from the medium treatment group chose her own life when $b_1 = 0.25$ and responded "I don't know" to $b_2 = 0.125$, she is counted as "wtp < 0.25" in Tables 2 and 4, but not in Table 3, as it is unclear whether her exact wtp is smaller or larger than 0.125.

[19] With a p-value of 0.0047 for the comparison between the low and the medium treatment group, and a p-value of 0.0105 for the comparison between the medium and the high treatment group.

Table 5 Frequencies of “alternative life” (a) or “own life” (o) responses by treatment

Treatment	Low	Medium	High
“o & o”	112 28.43%	106 27.89%	125 32.22%
“o & a”	19 4.82%	42 11.05%	43 11.08%
“a & o”	54 13.71%	47 12.37%	65 16.75%
“a & a”	180 45.69%	154 40.53%	110 28.35%
“I don’t know” (at least once)	29 7.36%	31 8.16%	45 11.60%
Total	394	380	388

the bids shown. Therefore, we compute the frequencies of each combination of responses by treatment group. If respondents were unbiased and answered truthfully, we expect large and highly significant differences between the treatments, as the corresponding bids differ strongly.

Table 5 displays the frequencies of “a” or “o” answers to both DC questions by treatment group, where “a” indicates choosing the alternative life and “o” indicates choosing the own life. Despite the large differences between the bids shown, the choices differ rather little. While respondents facing higher initial bids were significantly less likely to choose the alternative life twice, no clear pattern can be found for other combinations of answers.[20,21]

Hence, there is a strong indication that respondents did not answer completely at random and took the bids into account when making their choices, but were either biased, made some random errors, or did not answer truthfully. There may be other underlying mechanisms at work that also affect the DC responses, such as yea-saying or learning effects. To gain a clearer understanding, we analyse the responses to the OE questions.

3.2 Open-ended question

Next, we test whether the response behaviour in the OE question is independent of the initial bid. Since the OE questions were not restricted to the bounds set by the DC questions, respondents were free to state any $wtp < 1$.[22] Thus, the test of the OE responses helps to isolate the SPB from other biases and behavioural effect that may have affected the DC responses.

[20] Our results also indicate that respondents facing higher initial bids are less certain about their answers, as they answer “I don’t know” more often. However, this cannot be supported by a simple test using the self-assessed certainty about the answers; the correlation between self-assessed certainty and the initial bid only amounts to 0.0442 and is statistically insignificant. The five-point Likert scale from “completely certain” to “not certain at all” is mapped to the integers from 1 to 5 for this purpose. While this mapping function may be contestable (see Bond and Lang, 2019), it seems unlikely that a misspecification of the mapping function drives this result.

[21] The chi-squared test on the response pattern is highly significant (p-value < 0.000).

[22] In case their responses contradicted each other, respondents were asked if they wanted to correct their responses. Yet, even the corrected responses were often inconsistent (see Table 1). If respondents had been restricted to the interval stated in their DC answers, the initial bid would be strongly correlated with the WTP stated in the OE question, even if the responses to the OE question were entirely random.

Figure 3 Kernel densities of open-ended wtp by treatment

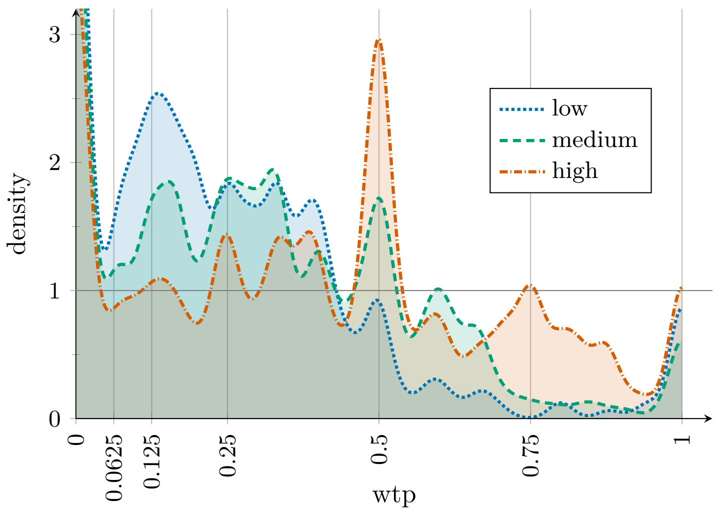


Figure 3 displays the kernel density estimation of responses to the OE question by treatment, where refusals and observations with wtp < 0 are filtered out.[23] There appears to be a clear relationship: the larger the first bid, the larger the expected response to the OE question. Especially towards the upper end of the wtp distribution (between 0.7 and 0.875), the probability mass for the high treatment group is still considerably large, but close to 0 for the low and medium treatment group. Conversely, the densities of the low and medium treatment groups are larger than those of the high treatment group in the range of low to medium values.

There are several peaks with varying sizes, depending on the treatment.[24] While less than 20% of respondents in the low treatment group state a wtp larger than 0.4, the same value lies at the median of the high treatment group. As expected, the average stated wtp of the medium treatment group lies between those of the other two groups. The correlation between the initial bid and stated wtp is 0.2685, which is considerable in size and highly significant (p -value < 0.000). This indicates that respondents are strongly influenced by the initial bid of the DC questions when giving an exact value to the OE question.

4 Parametric Analysis

Examining the densities by treatment alone does not provide a clear understanding of how the treatments affect the DC and OE responses specifically. These results are likely influenced by a combination of factors such as acquiescence bias, status quo bias, SPB, careless responding, and learning effects. Since the magnitude of these effects may vary among respondents, we conduct a parametric analysis to determine whether there is a systematic component in the heterogeneity

[23] Additionally, Table 11 in the Appendix contains some measures of the location and dispersion of these distributions.

[24] The first is located at wtp = 0 and is so large, that its top is cut off to improve readability. As shown in the last row of Table 11, a small but significant share of respondents state that they are not willing to forego any consumption in exchange for perfect health. Consistently with the findings above, this share is negatively correlated with the first bid. Other significant peaks in the density function are located at 0.125 for the low treatment group, at 0.5 for the medium and particularly for the high treatment group, at 0.75 for the high treatment group, and at 1 across all treatments. Interestingly, the densities at 0.25 and at 0.33 are roughly equal across all treatments, even though 0.33 was never shown as a bid and 0.25 was the most prominent bid. This indicates that respondents do not only use the bids, but also numbers that allow for easier calculations as anchors.

related to SPB. Additionally, we estimate the size of the SPB to facilitate a comparison with previous studies.

4.1 Treatment Effect

Our experimental design involves the random assignment of treatments. This ensures that the variable T_i, representing the treatment assignment for individual i, is independent of socio-economic characteristics. Leveraging this feature, we apply a simple linear model to estimate the effects of of the treatments on stated wtp. The equation to be estimated is the following:

$$\mathrm{wtp}_i = \beta_0 + \beta_T T_i + \beta_W W_i + \beta_S S_i + \beta_Q Q_i + \varepsilon_i, \tag{1}$$

where wtp_i denotes the relative WTP for perfect health of individual i, T_i is a vector of dummies indicating the treatment, W_i contains the well-being dimensions (health and consumption), S_i contains the socio-economic characteristics, Q_i encompasses survey-related data (e.g. evaluation of the questionnaire, survey completion time, etc.), and ε_i is a random error term. Our primary objective is to identify the effects of assigning a respondent to a specific treatment arm, which is captured by the vector β_T.[25] All remaining variables are only included as controls, helping to explain as much of the heterogeneity in stated wtp as possible. The main results of an ordinary least squares (OLS) estimation are presented in Table 6.

Both health and consumption (the well-being dimensions included in W_i), have a negative effect on wtp. This aligns with the expectation that individuals in poorer health are willing to pay more for perfect health, while having a higher level of consumption reduces the proportion of consumption an individual is willing to forego to improve health.

Now, we turn our attention to our primary variable of interest. In Model 1, we find the treatment dummies to be highly significant. The expected effect on wtp from being assigned to the low treatment, compared to the medium treatment, is a decrease of 0.0662 (meaning the proportion of their consumption they are willing to forego to obtain perfect health decreases by 6.62%). Conversely, the effect of assignment to the high treatment results in an expected increase of 9.87%. These results corroborate the hypothesis that the treatment had an effect on reported wtp.

We include a set of variables to control for the effect of the socio-economic characteristics S_i.[26] Marital status (being married) is found to have a negative effect on wtp. This could be attributed to the added challenges faced by single individuals in dealing with health issues without a support system.[27] Household size is found to positively influence wtp. In the questionnaire, respondents were asked to consider that reductions in their consumption would not impact that of their family or friends in the hypothetical scenario. Consequently, individuals benefiting from any kind of public good at their household level may perceive this as an indicator of increased availability of resources within their private sphere. Age was not found to have a significant effect.[28] Other socio-economic variables such as gender, parenthood status, nationality, occupation status, satisfaction with current health, and satisfaction with current consumption are not found to exhibit significant effects. Consequently, a more parsimonious model is preferred.[29] To reduce

[25] $\beta_T = \begin{pmatrix}\beta_l & \beta_h\end{pmatrix}'$ is a vector containing the parameters for the low and the high treatment group, while the medium treatment is the reference group.

[26] The complete list of estimation results is presented in the Appendix (Table ??).

[27] Additionally when partners share similar age and health levels, a radical increase in individual health may not be as desirable if it is not shared.

[28] Since preferences for health may vary with age, it is included in the models, despite its insignificant effect on wtp. The heteroscedasticity-robust F-test for the joint hypothesis H_0: $\beta_{\text{age}} = \beta_{\text{age}^2} = 0$ yields a p-value of 0.746.

[29] We collected information about political orientation, asking individuals to position themselves on a scale from 1 (extreme left) to 10 (extreme right). Interestingly, individuals identifying as right-wing exhibit a higher wtp for

Table 6 Linear regression of wtp on treatments

Dependent variable	Model 1 wtp	Model 2 wtp
Intercept	0.4803**	0.4459**
	(0.1560)	(0.1551)
Health	−0.0015***	−0.0015***
	(0.0004)	(0.0004)
Consumption (in logs)	−0.0224°	−0.0222°
	(0.0128)	(0.0125)
Low treatment	−0.0662***	−0.0525**
	(0.0168)	(0.0230)
× university	—	−0.0530
		(0.0343)
× familiar	—	0.0329
		(0.0426)
High treatment	0.0987***	0.1605***
	(0.0190)	(0.0260)
× university	—	−0.1186**
		(0.0406)
× familiar	—	−0.0906*
		(0.0455)
University	0.0001	0.0547*
	(0.0159)	(0.0260)
Familiar	−0.0068	0.0125
	(0.0187)	(0.0288)
Controls	Yes	Yes
R^2	0.132	0.148
Adj. R^2	0.119	0.131
Sample size	982	982
Estimation method	OLS	OLS

$^{***}p < 0.001$; $^{**}p < 0.025$; $^{*}p < 0.05$; $^{\circ}p < 0.1$. Std. errors in parentheses, calculated robust to heteroscedasticity.

the unexplained component we include a set of variables (vector Q_i) controlling for potential effects linked to the experience of the respondent answering the survey.[30]

To explore heterogeneity in the sensitivity to the bias, we interact the treatment variables with university education and familiarity with the hypothetical situation in Model 2. Interestingly, we observe some heterogeneity in the sensitivity to the treatments. For individuals with university education and those who report being able to imagine the hypothetical scenario clearly, we observe heterogeneous sensitivity to the bias induced by the high treatment. Respondents with a university education and those who can clearly imagine the hypothetical scenario are less affected by the high treatment.

If the induced bias was equal across groups, the SPB would be less relevant, as it would uniformly affect everyone. However, the fact that certain groups are biased more strongly than others implies that extra caution is necessary when comparing wtp across individuals from different groups. We will come back to this issue and discuss its implications for interpersonal well-being comparisons in Section 5. We also explore whether other characteristics are associated with heterogeneous sensitivity to the bias, finding no evidence of such patterns.[31] This is reassuring, as it suggests that while some heterogeneity in sensitivity to the bias exists, its average impact is limited in size. Nonetheless, we cannot exclude the possibility of heterogeneity in the bias sensitivity, for example for combinations of factors or heterogeneity in higher moments (such as the variance or skewness of the bias).

The regression results are robust to changes in the specification, including the inclusion or exclusion of additional variables, alternative variable definitions, or using uncorrected wtp as the dependent variable. However, statistically significant effects are observed less frequently in the latter case.[32]

4.2 Measuring Starting Point Bias

In the next step, we move to measuring the impact of the two bids. To do so, we implicitly need to assume that both bids have quantifiable effects on stated WTP which can be estimated separately from each other. The effect of the first bid would then represent the starting point bias. Most contributions in the SPB literature explicitly or implicitly assume that the follow-up bids have no effect on WTP; we intend to put this assumption to test. Let b_1 and b_2 denote the bids presented in the first and second DC questions, DC_1 and DC_2. An intuitive and straightforward method would be to regress wtp on b_1, b_2, and a set of controls. While our experimental design

perfect health. While the precise reason remains unclear, it might be attributed to a tendency among right-wing individuals to perceive health as a commodity tradeable within a market economy. However, since reporting this information was optional, there is a considerable amount of missing data (128 cases) for this variable. Given that this missingness is likely not at random, we opted to exclude this variable from the model and instead estimate our model with all individuals.

[30] We control for the time spent reading the initial screen, during which respondents were asked for their consent to participate. This time serves as an indicator of attentiveness. Additionally, we control for the variables "Evaluation: not clear" (indicating whether respondents found the questionnaire ambiguous), "Evaluation: biased" (reflecting perceptions of bias in the questionnaire), "Certainty" (measuring respondents' certainty about their answers), "Familiar" (indicating clarity in imagining the hypothetical situation), and "Corrections" (indicating whether respondents made changes to their OE responses after the initial submission). The evidence suggests that higher certainty in provided responses and not performing corrections to the initial statements increases wtp, aligning with the notion that individuals are more inclined to pay for something the more certain they are about what they are paying for.

[31] The list of variables tested includes: age, gender, political orientation, certainty, locus of control, satisfaction with health and consumption, being married, suffering from chronic diseases, low education, and evaluation of the questionnaire.

[32] A sensitivity analysis is presented in the Appendix (Table 15).

ensures the randomness of the treatment assignment, and therefore of b_1, this is not the case for b_2.

The bid presented in DC_2 is determined by the random value b_1 and the respondent's answer to DC_1, which depends on the prior "true" WTP.[33] Therefore, regressing wtp on b_2 is affected by endogeneity due to simultaneity. Note that this problem is inherent to any questionnaire involving a series of DC questions, where subsequent DC questions depend on the responses to previous ones.[34]

To address this problem, we employ an IV estimation strategy, utilising a novel set of instruments. We require the instrument(s)—denoted as Z—to causally determine b_2, and to affect wtp only through b_2 (Hansen, 2022, Chapter 12). If individuals answered the DC questions solely based on their prior wtp, instrumenting b_2 would be challenging, as its exogenous components would be accounted for entirely by b_1, a variable included in our model. However, in Section 3.1 we have shown evidence suggesting that other factors beyond wtp also drove the responses to DC_1. Our IV strategy aims to exploit these other mechanisms driving responses to DC_1 (such as assertiveness bias or display variations), to identify the exogenous effect of b_2 on wtp.

The set of instruments Z for b_2 are the following: a dummy variable indicating whether the individual responded to the questionnaire via laptop or mobile phone, a set of variables measuring the locus of control,[35] the number of clicks made, the time respondents spent looking at the screens displaying the DC questions, and a cubic transformation of the variable b_1.

Answering the questionnaire from a laptop or a mobile phone introduced random variation in how the options were displayed to respondents, and through this channel affected the likelihood of selecting the alternative life in the DC questions. Table 14 provides evidence from a logistic model predicting the probability of choosing the alternative life in the first DC question.[36] The type of device used was not found to have an effect on the reported wtp in the OE question, making it a suitable instrument. Table 13 displays the results of a regressions analysis where the set of instruments Z are utilised to explain both the endogenous and the dependent variables.

Regarding the time spent and the number of clicks made on the DC question screens, the mechanism is clear: the more time individuals spent on the DC questions, the more strongly their reported wtp was influenced by the bids. There are no apparent reasons why these variables would directly affect reported wtp except through the bids, making them suitable instruments.[37] Our results suggest that individuals with an internal locus of control are less likely to prefer the alternative life in the first DC question, which might be associated with these individuals being less prone to suffer from assertiveness bias.

[33] We use the term "true" WTP to denote the WTP a respondent would state in the absence of SPB. However, we do not claim that this "true" WTP is entirely unbiased, as virtually every aspect of the survey design and many other (potentially unknown) circumstances may systematically influence the respondents' decision. We do not even claim that an unbiased WTP exist; much in the spirit of the debate on whether preference elicitation should be viewed as an archaeological or an architectural exercise (see Slovic, 1995). Yet, the traditional concept serves its purpose in giving us an estimate for the impact of SPB.

[34] Depending on the exact survey design, randomising b_2 may result in showing respondents follow-up bids that can already be ruled out by the monotonicity assumption. This would not only be inefficient for estimating wtp, but could potentially confuse respondents.

[35] We incorporated four standardised questions to assess the locus of control. Participants were asked to express their level of agreement with the following statements: "I am my own boss", "If I work hard, I will have success", "In work and private life, what I do is mainly determined by others", and "Destiny often interferes in my plans". Respondents were categorised as having an internal or external locus of control based on their responses. For each of the variables, we also include a dummy variable indicating whether the individual completely agrees with the statement.

[36] No evidence was found that the device used correlates with the age, consumption, or health of respondents.

[37] As a sensitivity analysis we perform the IV analysis excluding the variables measuring the time spent and number of clicks in the second DC screen from Z, and we find consistent and qualitatively identical results.

Beyond the factors influencing responses to the first DC question, b_2 depends strongly on the treatment the respondent was assigned to. Using dummy variables to identify individual treatments is not feasible, as they would form a linear combination with b_1. Since our goal is to measure the effect of both bids, we cannot use b_1 (or any linear combination involving it) to instrument b_2. In Table 13 we observe that omitting the treatment assignment of individual i leads to weak explanatory power of the endogenous variable, risking the issue of weak instruments. To address this issue, we introduce a new variable that assigns values 1, 2 or 3 to each treatment assignment and then cubes the values. This transformation results in a variable that is not collinear with b_1, but has strong predictive power on b_2.

If we denote the exogenous variables as Z_1 and the instruments for b_2 as Z_2, we can test the rank condition by running a regression $b_{2,i} = \theta_1 Z_{1,i} + \theta_2 Z_{2,i} + e_i$ (Wooldridge, 2010, Chapter 5). We find strong evidence to reject the null hypothesis H_0: $\theta_2 = 0$.[38]

4.2.1 Linear Specification

To analyse the impact of the bids, we first use a linear model, similar to Equation 1. The equation to be estimated is:

$$\text{wtp}_i = \beta_0 + \beta_1 b_{1i} + \beta_2 b_{2i} + \beta_W W_i + \beta_S S_i + \beta_Q Q_i + \varepsilon_i. \tag{2}$$

Table 7 contains our main results.[39] In Model 3, both b_1 and b_2 are included. Here, the effect of b_1 is estimated to be close to zero (not statistically significant), while the effect of b_2 is strong (around 0.47) and highly significant (p-value < 0.000). The finding that only the second bid induces a bias but not the first goes against previous evidence, hinting at the endogeneity problem discussed above.

Model 4 applies our IV strategy for the endogenous variable b_2. The results align more closely with the existing literature, showing that the first bid has an effect of 0.34 (statistically significant at the 10% level), stronger than the effect of the second bid, which is estimated to be around 0.10 and statistically insignificant. However, the standard errors are large. Consequently, omitting b_2 might provide a reasonable estimation of the strength of the SPB.[40] Therefore, we exclude the second bid in Model 5, yielding an estimated effect of b_1 of around 0.43. Nonetheless, this model might suffer from omitted variable bias due to the exclusion of b_2; hence, our estimate captures the combined effect of the bids in our experimental setting and is likely larger than the starting point bias.

4.2.2 Convex Specification

Herriges and Shogren (1996) have proposed to model the mechanism of the SPB as a convex combination between the prior WTP_{0i} that individuals hold before seeing any DC question, and the bid presented to them (b_i):

$$\text{wtp}_i = (1 - \gamma)\, \text{wtp}_{0i} + \gamma\, b_i + \varepsilon_i. \tag{3}$$

[38] Performing a Wald test to account for heteroscedasticity we reject the null hypothesis H_0: $\theta_2 = 0$ (p-value $<$ 0.000).

[39] The effects of the variables included in the vectors S_i and Q_i are omitted for brevity. Nonetheless, the effects observed are consistent with the estimated effects discussed in the previous section. The complete list of estimators can be found in the Appendix (Table **??**). Additionally, a sensitivity analysis is provided in Table 16 of the Appendix.

[40] As described in the previous section, we instrument b_2 using the set of instruments Z_2. The IV diagnostics reject that we have weak instruments (p-value < 0.0001), the Wu-Hausman test supports the need to use an IV strategy (p-value $= 0.0249$), and the Sargan over-identification statistic has a p-value of 0.7092, indicating that there is no evidence that the model is over-identified. If we opt against including the cube of the treatment variable as an instrument in Z_2, the p-values for these statistics are 0.0385, 0.0633, and 0.8749 respectively, indicating weaker support for the IV strategy.

Table 7 Linear regression of wtp on bids

Dependent variable	Model 3 wtp	Model 4 wtp	Model 5 wtp
Intercept	0.2688°	0.3400^{*}	0.3591^{**}
	(0.1459)	(0.1553)	(0.1545)
Health	-0.0006°	-0.0013^{**}	-0.0015^{***}
	(0.0003)	(0.0005)	(0.0004)
Consumption (in logs)	-0.0239^{*}	-0.0224°	-0.0221°
	(0.0121)	(0.0124)	(0.0126)
b_1	0.0101	0.3434°	0.4328^{***}
	(0.0593)	(0.1821)	(0.0499)
b_2	0.4673^{***}	0.0988	—
	(0.0414)	(0.1919)	
Controls	Yes	Yes	Yes
R^2	0.247	0.175	0.132
Adj. R^2	0.236	0.163	0.119
Sample size	982	982	982
Estimation method	OLS	IV	OLS

$^{***}p < 0.001$; $^{**}p < 0.025$; $^{*}p < 0.05$; $^{\circ}p < 0.1$. Std. errors in parentheses, calculated robust to heteroscedasticity.

The bid presented to respondents in the DC acts as an anchor, causing wtp_i to approach b_i. The strength of the SPB is measured by the parameter γ, capturing the anchoring effect of b_i (with $0 \leq \gamma \leq 1$). Assuming that the true wtp_{0i} can be explained by a vector of socio-economic variables X_i plus a random error term ϵ_i, we can rewrite the expression above to obtain:[41]

$$\text{wtp}_i = (1-\gamma)\,\beta_X\,X_i + \gamma\,b_{1i} + \varepsilon'_i, \tag{4}$$

where $\varepsilon'_i = \varepsilon_i + (1-\gamma)\,\epsilon_i$, and β and γ are parameters to be estimated.

Notably, Herriges and Shogren (1996) examine the influence of the initial bid on wtp, excluding subsequent bids from their model, an extension discussed but not empirically tested in their paper. Nolte (Lechner), Rozan, and Laisney (2006) have extended their model to incorporate the impact of the second bid. However, they assume that both the first and the second bid equally affect wtp, an assumption they acknowledge requires further testing in future research.

Since we want to allow for different effect sizes of b_1 and b_2, we opt for an approach of sequentially updated wtp with two bias parameters. Let wtp_1 denote the (latent) wtp updated after seeing the first bid (b_1), and wtp_2 the reported wtp updated after seeing the second bid (b_2), respectively:

$$\begin{aligned}
\text{wtp}_{0,i} &= \beta_X\,X_i + \varepsilon_{0,i}\\
\text{wtp}_{1,i} &= (1-\gamma)\,\text{wtp}_{0,i} + \gamma\,b_{1,i} + \eta_{1,i}\\
\text{wtp}_{2,i} &= (1-\omega)\,\text{wtp}_{1,i} + \omega\,b_{2,i} + \eta_{2,i}.
\end{aligned}$$

[41] The vector X_i includes the well-being dimensions and socio-economic variables that influence wtp. In other words, the vector X_i encompasses the information from vectors W_i and S_i as specified in Equation 1 and Equation 2.

Table 8 Non-linear regression of wtp on bids

Dependent variable	Model 6 wtp	Model 7 wtp	Model 8 wtp
Intercept	1.7348***	0.6518**	0.8487***
	(0.3151)	(0.2707)	(0.2201)
Health	−0.0020**	−0.0025**	−0.0021***
	(0.0007)	(0.0008)	(0.0005)
Consumption	−0.0920***	−0.0406°	−0.0439**
(in logs)	(0.0261)	(0.0220)	(0.0182)
b_1	0.2450**	0.4211***	0.2963***
	(0.0826)	(0.0990)	(0.0498)
b_2	0.3329***	0.0290	—
	(0.0422)	(0.1676)	
Controls	Yes	Yes	Yes
J-test statistics	—	7.4341	—
J-test p-value	—	0.9167	—
Sample size	982	982	982
Estimation method	Iter. GMM	Iter. GMM (IV)	Iter. GMM

$^{***}p < 0.001$; $^{**}p < 0.025$; $^{*}p < 0.05$; $^{\circ}p < 0.1$. Std. errors in parentheses, calculated robust to heteroscedasticity.

By combining these expressions, we derive our model of interest:

$$\text{wtp}_i = (1 - \gamma - \omega + \gamma\,\omega)\,\beta_X\,X_i + \gamma\,(1 - \omega)\,b_{1i} + \omega\,b_{2i} + \eta_i^*, \tag{5}$$

where γ accounts for the bias induced by the first bid, ω for the second bid, and η_i^* is a residual term.[42] Due to the multiplicative nature of some parameters in the model, we prefer an estimation method that can separately identify each parameter.[43] To account for the endogeneity of b_2 we instrument this variable using the instruments outlined in the previous sections. Equation 5 is estimated using the GMM, implemented in R using the `momentfit` package.[44] The main results are presented in Table 8.[45]

Mirroring the logic followed in Section 4.2.1, Model 6 in Table 8 regresses on both b_1 and b_2, not accounting for endogeneity. Model 7 applies our IV strategy to instrument b_2, while Model 8 estimates a constrained version of Equation 5 where $\omega = 0$ is imposed. The results corroborate our findings outlined in Section 4.2.1. In Model 6, the outcomes again appear counter-intuitive: the effect of b_2 is estimated to be stronger than b_1. Employing an IV strategy, in Model 7, we

[42] The term $\varepsilon_{0,i}$ represents a random variable following a normal distribution with a mean of zero. The terms $\eta_{1,i}$ and $\eta_{2,i}$ are residual components, each comprising two parts: an independent random error term, and a deterministic component that captures survey-specific features, analogous to the vector Q_i in Equation 1 and Equation 2. By combining these expressions we derive the residual term η_i^* in Equation 5. A detailed explanation is provided in the Appendix (Section A.5).

[43] Consider estimating Equation 5 using OLS. One would obtain an estimate for $\gamma^* = \gamma\,(1 - \omega)$, requiring the calculation $\widehat{\gamma} = \frac{\widehat{\gamma}^*}{1-\widehat{\omega}}$ to obtain an estimate for γ. This operation involves a ratio between estimators. Identifying β involves an even more complex ratio among several estimated parameters. Therefore, an estimation method able to provide a direct estimation of the parameters of interest—such as non-linear least squares or GMM—is preferable.

[44] The `momentfit` package (Chaussé, 2023) is an updated version of the `gmm` package (see also Chaussé, 2021).

[45] The complete output can be found in the Appendix (Table 12).

estimate $\widehat{\gamma} = 0.4211$, with the second bid showing a small effect that lacks statistical significance. In Model 8, where $\omega = 0$ is imposed, we observe that the effect of γ is approximately 0.2963 and more precisely estimated.

Aprahamian, Chanel, and Luchini (2007) use a similar convex specification accounting for heterogeneity in the bias. They provide simulations showing that if the bias is heterogeneous but independent of the covariates X_i, parameter estimates remain unbiased—and therefore the predicted value of wtp as well—but have inflated standard errors, impeding inference. However, if the heterogeneous bias is related to the covariates X_i, the estimation of the parameters is biased. In their model, they assume that only the first bid (and not the second) exerts an effect on wtp.

We incorporate heterogeneity in the bias generated by b_1 to our model by changing the parameter γ to $\gamma' = (\gamma + \lambda H_i)$ in Equation 5, assuming the effect of b_2 is homogeneous but distinct from the effect of b_1. We explore which characteristics in H_i are related to the effect of b_1 on wtp (captured by the vector of parameters λ). The variables tested include health, consumption, gender, education (whether or not the respondent has received tertiary education), age, locus of control, evaluation of the questionnaire, and reported certainty with the answers; yet, we do not find statistically significant evidence that any of these characteristics is linked to heterogeneity in the bias sensitivity. Therefore, we prefer a model with homogeneous bias.[46]

As previously mentioned, earlier estimates of SPB have focused on measuring WTP for environmental protection. Herriges and Shogren (1996) report $\gamma = 0.36$ for one part of the sample (recreationists), and no evidence of SPB for the other part of the sample (local residents). Nolte (Lechner), Rozan, and Laisney (2006) assume consistent anchoring effects across the first and second bids and report $\gamma = 0.63$. Whitehead (2002) reports $\gamma = 0.54$ when incorporating a structural shift for incentive incompatibility. Chien, Huang, and Shaw (2005) use a double-bounded model distinguishing anchoring from yea-saying and report $\gamma = 0.65$. Since each valued good is unique across studies, it is challenging to compare them. Our estimator for SPB is smaller than in (most of) the previous studies, which might be explained by the fact that we are studying this phenomenon in a different domain (health); this aligns with the notion that individuals are less subject to cognitive biases when valuing goods for which they have more clearly defined preferences.

5 Equivalent Consumption

After confirming the presence of the SPB and giving an estimate of its relative size, we now present two methods to correct for the bias. We construct a well-being ranking of all individuals using the equivalent consumption measure and compare the results with and without corrections. Equivalent consumption is a preference-based well-being measure which denotes the level of consumption an individual has left after subtracting her WTP for a reference level in all non-monetary life dimensions, in our case for perfect health. This way we capture two dimensions of well-being in one single measure while respecting that individuals may have different ideas about what constitutes a good life.[47]

Here, we use the stated WTP for perfect health from the OE question to compute equivalent consumption for each individual: $c_i^* = c_i - \mathrm{WTP}_i$. If an individual i has a higher equivalent

[46] A word of caution is in order regarding our approach to model heterogeneity in the bias. Veronesi, Alberini, and Cooper (2011) provide evidence that using a set of dummy variables to account for SPB has low power, which might explain why we do not find evidence to reject $\lambda = 0$. Nonetheless, in the specification where heterogeneity is modelled, the estimates for other model parameters remain consistent—despite the insignificance of λ—suggesting unbiased results when not modelling bias heterogeneity.

[47] For more detailed explanations and an axiomatic characterisation of the equivalent consumption measure, see Decancq, Fleurbaey, and Schokkaert (2015) and Fleurbaey (2016).

consumption than another individual j, we conclude that the well-being of individual i is higher than j's. However, since we showed that i's and j's response behaviour were influenced by the starting point, such a ranking may not be stable; if the respondents had seen different starting points, their well-being ranking could be reversed. In the following part, we explore the extent of this problem. Note that this is an illustration in which representativeness and imprecision are disregarded.

In contrast to Liou (2015), who calculates the prior WTP by estimating the SPB and inverting the equivalent of our Equation 3, we do not assume that the prior WTP in Equation 3 is equal to true WTP, safe that true WTP even exists.[48] Instead, we view every starting point—i.e. each experimental treatment—as a frame under which a choice was made. Since we only observe the responses under the actual frame of choice to which individuals were assigned, we impute hypothetical responses for each individual under every other frame.

We propose two approaches for imputing hypothetical equivalent consumption values, a non-parametric and a parametric approach. While the former minimises the required assumptions and adheres closely to stated WTP, the latter adjusts stated preferences to align with average preferences, thereby correcting for potential errors in respondents' stated WTP.

The actual and hypothetical values are then ranked by the two different methods. In the first, we rank individuals by the frame that yields their lowest equivalent consumption, which need not be the same for each individual. This acts as a prioritarian safety principle and resembles the safety principle proposed in Fleurbaey and Schokkaert (2013).[49] Second, we compare all frames at once, ranking one individual as better off than another only if her (hypothetical) equivalent consumption is larger under each frame. Since one individual may be better off under some, but not all frames, this criterion may yield incomplete rankings.

5.1 Non-parametric correction

As our experimental design consists of a small number of treatments, it is possible to impute hypothetical equivalent consumption values without imposing parametric assumptions on WTP. Assuming that the assignment to the treatment group is independent of the true WTP—which is satisfied by design—and that stated WTP is a monotonic function of true WTP, it follows that respondent i's relative position in the cumulative distribution function (CDF) of stated WTP is expected to be equal across treatments. Thus, i's hypothetical WTP under a different treatment can be imputed by reading off the quantile in the stated WTP distribution of the different treatments. Figure 8 visualises this idea: in the empirical CDF of the three treatments, we can impute i's hypothetical WTP by first determining her position on the vertical axis in her own treatment, and then reading off the value at another treatment on the same horizontal line.[50]

The results of ranking each pair of respondents using these methods to compute the hypothetical equivalent consumption are summarised in the first part of Table 9. It shows the percentage of ranking reversals when changing from the equivalent consumption computed using the actual response to the equivalent consumption using the other two approaches, as well as the percentage of incompleteness, i.e. the percentage of comparisons where no respondent can be identified as better off.

[48] We refrain from taking this assumption since the elicitation method itself may have an impact on stated WTP. Thus, it can ultimately never be verified which elicitation yields the true WTP, which in turn opens up the question of whether the concept of true WTP has any empirical value, and WTP should always be interpreted in the context it was measured.

[49] The safety and super safety principle in Fleurbaey and Schokkaert (2013) are specifically designed to allow for rankings of equivalent consumption when preferences are context-dependent.

[50] A limitation with this approach is that the CDF of stated WTP are usually not continuous, as can be seen in several large jumps (for instance, at 0.33 and 0.5).

Table 9 Reversals and incompleteness by ranking method (in %)

Sample	Full	By consumption decile									
Consumption decile		1	2	3	4	5	6	7	8	9	10
Non-parametric correction											
Ranking reversals											
Actual vs. lowest	7.57	11.71	9.31	7.72	10.99	7.80	7.00	10.45	9.05	9.26	8.06
Actual vs. all frames	2.30	4.66	6.09	5.07	5.97	5.44	4.75	7.10	3.81	6.34	3.47
Incompleteness											
Actual	1.00	3.98	4.09	4.48	3.76	4.07	4.68	3.18	2.66	3.53	1.71
Lowest	0.40	2.10	2.47	2.74	1.64	1.43	2.44	1.04	1.57	0.59	1.09
All frames	9.43	11.71	8.60	7.68	8.82	6.83	7.07	7.07	8.97	7.83	9.74
Parametric correction											
Ranking reversals											
Predicted vs. lowest	6.22	18.77	26.99	33.54	27.62	33.00	30.59	30.63	24.74	28.56	12.92
Predicted vs. all frames	3.19	9.06	13.00	15.24	11.42	14.44	11.79	14.60	8.33	11.95	6.66
Incompleteness											
Predicted	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Lowest	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
All frames	5.73	17.26	26.87	34.06	29.38	32.75	32.71	33.84	32.79	32.01	13.76

Since people tend to round numbers when responding to questions about quantities, many respondents in our questionnaire have the same stated consumption and WTP. The consequences of this rounding heuristic is reflected by the incompleteness of rankings using the uncorrected stated preferences (see the third row of Table 9). We observe that the level of incompleteness increases when comparing equivalent consumption at all frames. This is to be expected, as the ranking at all frames is the least discerning ranking, meaning that it is most likely to yield incomplete rankings. Note that individuals within the same quantile of the CDF in their respective treatments exhibit the same wtp across all frames.

Turning to the question of ranking reversals, a small but notable percentage of rankings is reversed when comparing actual with corrected equivalent consumption measures. This percentage of reversals amounts to 7.57% when comparing the actual ranking with the ranking at the lowest frame, and to 2.30% when comparing it with the ranking at all frames. However, a ranking between two individuals is unlikely to be reversed if the difference in their consumption level is too big, regardless of their WTP or treatment. To understand the role of WTP among individuals with similar consumption levels better, we also compare individuals within consumption deciles.[51] When comparing the actual frame to the lowest frame, the percentage of reversals ranges from 7.00% in the sixth consumption decile to 11.71% in the first decile, with no clear pattern across deciles. The percentage of reversals between the actual ranking and the ranking at all frames is on a lower level, ranging from 3.47% to 7.10%.[52]

5.2 Parametric correction

Another possibility to correct for SPB is to use the parametric model presented in Section 4 to predict WTP for each treatment. For brevity, and to use the least number of parametric assumptions, we use Model 2, which relates WTP to the treatments and allows for heterogeneity in the bias sensitivity (see Equation 1 and the second column of Table 6). Figure 9 shows the empirical CDF of predicted WTP by treatment.[53] It differs from Figure 8 in two aspects:[54] first, the CDF appears to be continuous, and second, the volatility is much smaller.

The percentages of incompleteness by approach and ranking reversals between methods for the parametric approach can be found in the second part of Table 9. The levels of incompleteness obtained with the parametric approach differ strongly from those obtained with the non-parametric approach. All rankings are defined when comparing predicted equivalent consumption at the actual frames, or when using the lowest frame[55] For the rankings under all frames, 5.73% of all possible rankings are not feasible. Interestingly, when we examine incompleteness among individuals within the same consumption deciles, we observe high levels of incompleteness, reaching up to 34% for some deciles. This occurs because individuals with similar consumption

[51] Since respondents often corrected their consumption value to multiples of 100, a random tie-breaker is used to determine consumption deciles of equal sizes. The percentages of ranking reversals are simulated by running this randomisation 10 000 times.

[52] The percentage of reversals between the ranking at the lowest value and the ranking at all frames is not shown, as the two methods cannot yield opposite rankings. To see why this holds, consider two individuals where the ranking at the lowest frame considers one individual better off than another. The individual who is worse off has the lowest (hypothetical) equivalent consumption level; this level cannot be larger than any (hypothetical) equivalent consumption level of the other individual.

[53] Since WTP under the actual treatment is a predicted value in the parametric approach, it is generally not equal to stated WTP.

[54] Note that the two methods can be compared directly because we restrict the analysis to the same sample of respondents.

[55] This can be attributed to the (quasi-)continuity of the CDF of predicted WTP, which ensures that two individuals have a very low chance of having the same predicted equivalent consumption. Further, we do not observe individuals with the same set of predictors.

levels but varying sensitivity to the bias may exhibit higher wtp under one frame but not under another. This finding is noteworthy as it demonstrates that the frame, i.e. the starting point used in the DC question, is not neutral. In other words, when comparing individuals with similar levels of consumption but different sensitivity to the bias, the choice of the frame can reverse their rankings.

Within every decile and for every method, the percentage of reversals compared to the ranking using the actual frames is higher for the parametric than for the non-parametric correction. In the first and tenth decile, the levels are still moderate, but particularly from the second to the seventh decile, the percentage of reversals is sizeable. When comparing at the lowest frame, the percentage of reversals lies above 20% everywhere from the second to the ninth decile, ranging up to almost 34%. For the approach comparing all frames at once, the percentage of reversals is on a lower, but yet considerable level. These findings confirm that the correction for SPB has a relevant effect on ranking individual well-being using the equivalent consumption approach, particularly for individuals with a similar consumption level.

6 Conclusion

This paper investigates the effect of the starting point bias on stated willingness to pay. While its presence in the environmental economics literature, where respondents are asked for their willingness to pay for protecting lakes, clean air, etc., is well-established, there is only little evidence for less abstract goods, such as one's own health. Further, very little work has been done on developing a framework to correct for the bias. To close these gaps, we conducted an online questionnaire that asked respondents to state their willingness to pay for perfect health. The respondents first answered two dichotomous choice questions, in which they could indicate whether they preferred an alternative life (with perfect health and a lower level of consumption) or their own life. Afterwards, they were asked to state a monetary value at which they would be indifferent between the two lives. The experimental treatment consisted of showing different bids in the first dichotomous choice question (low, medium, or high).

Our tests to verify the presence of the bias and our estimation of the bias size leverage several methodological approaches, including non-parametric and parametric techniques. In a nutshell, we find robust evidence confirming starting point bias in the context of willingness to pay estimation for health. While comparing estimates across studies can be challenging, our results suggest that the size of starting point bias in our context is smaller than in most of the previous studies focusing on the context of willingness to pay for environmental protection.

A non-parametric analysis indicates that the open-ended responses show strong sensitivity to the starting point, while the dichotomous choice responses may be influenced by a broader range of factors. Although our data does not allow us to isolate them, these findings highlight the need for careful consideration when relying solely on dichotomous choice questions to estimate willingness to pay.

The fact that responses to dichotomous choice questions seem to be influenced by other factors besides willingness to pay provides us with an opportunity to use a novel instrumental variable strategy to estimate the effect of the second bid. Our estimate for the influence of the second bid on stated willingness to pay is not significantly different from zero and likely smaller than that of the initial bid, suggesting that the initial bid is the main driver in triggering the observed bias. Therefore, our findings support the assumptions stated in previous studies, which, to the best of our knowledge, have not been empirically tested before.

The parametric analysis further allows for modelling heterogeneity in the sensitivity to starting point bias for different groups. While there is no evidence for heterogeneity when estimating the effect of both bids, we find significant interactions in simpler models that merge the effect of the bids using treatment dummies. The data suggest that individuals with university education and those who perceive the imaginary scenario as familiar are less sensitive to high starting bids. These differences in sensitivity are large enough to offset the bias of seeing a high starting point for the group of individuals with university education and familiarity with the scenario.

Using the estimation results, we present a non-parametric and a parametric approaches to correct for starting point bias. As an illustration, we apply the approaches on the equivalent consumption measure, a preference-based well-being measure, and check the extent of re-ranking compared to the uncorrected estimates. Since equivalent consumption is strongly correlated with (regular) consumption, the extent of re-ranking in the full sample is small. However, when focusing on rankings between individuals from the same consumption decile, the percentage of ranking reversals is sizeable. This is particularly the case for the parametric approach, which removes most of the individual variation and thus foregrounds the role of heterogeneity in the bias sensitivity.

Our findings should caution researchers when selecting the initial bids in designing a dichotomous choice questionnaire for willingness to pay estimation. Generalising from our study, we recommend choosing a starting point close to the expected median, and to include an open-ended follow-up question, as responses to dichotomous choices seem to be influenced by factors other than willingness to pay.

Yet, it is important to note that our data was collected through an online survey. Future research could test these conclusions in more controlled settings. The issue of willingness to pay elicitation through stated preference techniques being valid has been extensively debated in the literature for decades. While some researchers argue that behavioural biases can be minimised with careful experimental designs, others contend that such recommendations have not been empirically validated. The recent development of new techniques to monitor and understand heuristic processes in controlled environments—such as laboratories—offers a promising avenue to address some of these longstanding questions.

References

Alberini, Anna, Barbara Kanninen, and Richard T. Carson (1997). “Modeling Response Incentive Effects in Dichotomous Choice Contingent Valuation Data”. In: *Land Economics* 73.3, pp. 309–324.

Aprahamian, Frédéric, Olivier Chanel, and Stéphane Luchini (2007). “Modeling Starting Point Bias as Unobserved Heterogeneity in Contingent Valuation Surveys: An Application to Air Pollution”. In: *American Journal of Agricultural Economics* 89.2, pp. 533–547.

Araña, Jorge E. and Carmelo J. León (2007). “Repeated Dichotomous Choice Formats for Elicitation of Willingness to Pay: Simultaneous Estimation and Anchoring Effect”. In: *Environmental and Resource Economics* 36.4, pp. 475–497.

Ariely, Dan, George Loewenstein, and Drazen Prelec (2003). “’Coherent Arbitrariness’: Stable Demand Curves Without Stable Preferences”. In: *The Quarterly Journal of Economics* 118.1, pp. 73–106.

Arrow, Kenneth, Robert Solow, Paul R. Portney, Edward E. Leamer, Roy Radner, and Howard Schuman (1993). *Report of the NOAA Panel on Contingent Valuation*.

Bernheim, B. Douglas and Antonio Rangel (2009). “Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics”. In: *The Quarterly Journal of Economics* 124.1, pp. 51–104.

Bond, Timothy N. and Kevin Lang (2019). “The Sad Truth about Happiness Scales”. In: *Journal of Political Economy* 127.4, pp. 1629–1640.
Boyle, Kevin J. (1990). “Dichotomous-Choice, Contingent-Valuation Questions: Functional Form Is Important”. In: *Northeastern Journal of Agricultural and Resource Economics* 19.2, pp. 125–131.
Boyle, Kevin J., Richard C. Bishop, and Michael P. Welsh (1985). “Starting Point Bias in Contingent Valuation Bidding Games”. In: *Land Economics* 61.2, pp. 188–194.
Brookshire, David S., Berry C. Ives, and William D. Schulze (1976). “The Valuation of Aesthetic and Preferences”. In: *Jornal of Environmental Economics and Management* 6, pp. 325–346.
Cameron, Trudy Ann and John Quiggin (1994). “Estimation Using Contingent Valuation Data from a ’Dichotomous Choice with Follow-Up’ Questionnaire”. In: *Journal of Environmental Economics and Management* 27, pp. 218–234.
Chaussé, Pierre (2021). “Computing generalized method of moments and generalized empirical likelihood with R”. In: *Journal of Statistical Software* R package version 0.5.34, pp. 1–35.
— (2023). *momentfit: Methods of Moments*. R package version 0.5.
Chien, Yu-Lan, Cliff J. Huang, and Daigee Shaw (2005). “A general model of starting point bias in double-bounded dichotomous contingent valuation surveys”. In: *Journal of Environmental Economics and Management* 50.2, pp. 362–377.
Cooper, Joseph C. (2002). “Flexible Functional Form Estimation of Willingness to Pay Using Dichotomous Choice Data”. In: *Journal of Environmental Economics and Management* 43.2, pp. 267–279.
Da Costa, Shaun, Koen Decancq, Marc Fleurbaey, and Erik Schokkaert (2024). *Preference elicitation methods and equivalent income: an overview*. Tech. rep. 24/09. Herman Deleeck Centre for Social Policy.
Decancq, Koen, Marc Fleurbaey, and Erik Schokkaert (2015). “Happiness, Equivalent Incomes and Respect for Individual Preferences”. In: *Economica* 82.5, pp. 1082–1106.
Decancq, Koen and Annemie Nys (2021). “Non-parametric well-being comparisons”. In: *European Economic Review* 133, pp. 1–21.
Fleurbaey, Marc (2016). “Equivalent Income”. In: *The Oxford Handbook of Well-Being and Public Policy*. Ed. by Matthew D. Adler and Marc Fleurbaey. Vol. 16. Oxford University Press, pp. 453–475.
Fleurbaey, Marc and Erik Schokkaert (2013). “Behavioral Welfare Economics and Redistribution”. In: *American Economic Journal: Microeconomics* 5.3, pp. 180–205.
Hammack, Judd and Gardner Mallard Brown Jr. (1974). *Waterfowl and Wetlands: Toward Bioeconomic Analysis*. Johns Hopkins University Press.
Hanemann, W. Michael and Barbara Kanninen (1996). *The Statistical Analysis of Discrete-Response CV Data*. CUDARE Working Papers 798. Department of Agricultural and Resource Economics.
Hansen, Bruce E. (2022). *Econometrics*. Princeton: Princeton University Press. 1044 pp.
Heinzen, Rebekah R. and John F. P. Bridges (2008). “Comparison of four contingent valuation methods to estimate the economic value of a pneumococcal vaccine in Bangladesh”. In: *International Journal of Technology Assessment in Health Care* 24.04, pp. 481–487.
Herriges, Joseph A. and Jason F. Shogren (1996). “Starting Point Bias in Dichotomous Choice Valuation with Follow-Up Questioning”. In: *Journal of Environmental Economics and Management* 30, pp. 112–131.
Hole, Arne Risa (2007). “A Comparison of Approaches to Estimating Confidence Intervals for Willingness to Pay Measures”. In: *Health Economics* 16.8, pp. 827–840.

León, Roberto and Carmelo J. León (2003). "Single or Double Bounded Contingent Valuation? A Bayesian Test". In: *Scottish Journal of Political Economy* 50.2, pp. 174–188.

Li, Chuan-Zhong and Leif Mattsson (1995). "Discrete Choice under Preference Uncertainty: An Improved Structural Model for Contingent Valuation". In: *Journal of Environmental Economics and Management* 28.2, pp. 256–269.

Liou, Je-Liang (2015). "Bias correcting model of starting point bias with censored data on contingent valuation method". In: *Environmental Economics* 6.3, pp. 8–14.

Nolte (Lechner), Sandra, Anne Rozan, and Francois Laisney (2006). *A Model of the Anchoring Effect in Dichotomous Choice Valuation with Follow-Up*. Working Papers 2003-07. University of Strasbourg.

Onwujekwe, Obinna and Douglas Nwagbo (2002). "Investigating Starting-Point Bias: A Survey of Willingness to Pay for Insecticide-Treated Nets". In: *Social Science & Medicine* 55, pp. 2121–2130.

Polome, Philippe, Anne van der Veen, and Peter Geurts (2006). "Is Referendum the Same as Dichotomous Choice Contingent Valuation?" In: *Land Economics* 82.2, pp. 174–188.

Randall, Alan, Orlen Grunewald, Sue Johnson, Richard Ausness, and Angelos Pagoulatos (1978). "Reclaiming Coal Surface Mines in Central Appalachia: A Case Study of the Benefits and Costs". In: *Land Economics* 54.4, pp. 472–489.

Randall, Alan, Berry C. Ives, and Clyde Eastman (1974). "Bidding games for valuation of aesthetic environmental improvements". In: *Journal of Environmental Economics and Management* 1.2, pp. 132–149.

Rowe, Robert D., Ralph C. d'Arge, and David S. Brookshire (1980). "An Experiment on the Economic Value of Visibility". In: *Journal of Environmental Economics and Management* 7, pp. 1–19.

Salant, Yuval and Ariel Rubinstein (2008). "(A,f): Choice with Frames". In: *Review of Economic Studies* 75, pp. 1287–1296.

Samples, Karl C. (1985). "A Note on the Existence of Starting Point Bias in Iterative Bidding Games". In: *Western Journal of Agricultural Economics* 10.1, pp. 32–40.

Slovic, Paul (1995). "The Construction of Preference". In: *American Psychologist* 50.5, pp. 364–371.

Stiglitz, Joseph, Amartya Sen, and Jean Fitoussi (2009). *Report of the Commission on the Measurement of Economic Performance and Social Progress (CMEPSP)*. Tech. rep. Commission on the Measurement of Economic Performance and Social Progress.

Thayer, Mark A. (1981). "Contingent Valuation Techniques for Assessing Environmental Impacts: Further Evidence". In: *Journal of Environmental Economics and Management* 8, pp. 27–44.

Tversky, Amos and Daniel Kahneman (1974). "Judgment under Uncertainty: Heuristics and Biases". In: *Science* 185.4157, pp. 1124–1131.

Veronesi, Marcella, Anna Alberini, and Joseph C. Cooper (2011). "Implications of Bid Design and Willingness-To-Pay Distribution for Starting Point Bias in Double-Bounded Dichotomous Choice Contingent Valuation Surveys". In: *Environmental and Resource Economics* 49.2, pp. 199–215.

Whitehead, John C. (2002). "Incentive Incompatibility and Starting-Point Bias in Iterative Valuation Questions". In: *Land Economics* 78.2, pp. 285–297.

Wooldridge, Jeffrey M. (2010). *Econometric analysis of cross section and panel data*. Second edition. Cambridge, Massachusetts: MIT Press. 1974 pp.

A Appendix

A.1 Example questions

Figure 4 First DC question, example for $c_i = 1200$ in low treatment

Imagine que es posible optar por vivir una **vida alternativa** durante el próximo año, con absolutamente **ningún problema ni complicación de salud**, pero en contrapartida, debe **reducir su nivel de consumo mensual** (recuerde no tomar en cuenta el alquiler/hipoteca en el consumo).

	Su Vida	**Vida alternativa**
Consumo	1200€	1050€
Salud	salud actual	salud perfecta

Considere las dos opciones presentadas en la tabla.

Puede escoger entre la **vida alternativa**, con **perfecta salud y 1050€** por mes para su consumo, y **su vida actual**. Asuma que todos los otros aspectos de su vida permanecen inalterados, únicamente su consumo personal (no el de su familia o amigos) se ve afectado.

¿Qué vida escogería?

Mi vida actual | **La vida alternativa** | **No lo sé**

Translation of Figure 4:

Translation of Figure 4:

Imagine it was possible for you to live an **alternative life** for the next year, with **no health problems or complications at all**, but in return you would have to **reduce your monthly level of consumption** (remember not to take your rent/mortgage payments into account for consumption).

	Your life	**Alternative life**
Consumption	1200€	1050€
Health	actual health	perfect health

Consider the two options presented in the table.

You can choose between the **alternative life**, with **perfect health and 1050€** per month for your consumption, and **your actual life**. Assume that all other aspects of your life remain unchanged, only your personal consumption (not that of your family or friends) is affected.

Which life would you choose?

My actual life	**The alternative life**	**I don't know**
○	○	○

Figure 5 Second DC question, example for $c_i = 1200$ in low treatment

Ha indicado que preferiría la **situación alternativa** (con **perfecta salud y 1050€** por mes para consumo) sobre **su situación actual** (con **su salud actual y 1200€** por mes para consumo).

En la siguiente tabla, le presentamos otra elección, dónde hemos **modificado el nivel de consumo** en la vida alternativa.

	Su Vida	Vida Alternativa
Consumo	1200€	900€
Salud	salud actual	salud perfecta

Considere las dos opciones presentadas en la tabla.

Puede escoger entre la **vida alternativa**, con **perfecta salud y 900€** por mes para su consumo, y **su vida actual**. Nuevamente, asuma que todos los otros aspectos de su vida permanecen inalterados, únicamente su consumo personal (no el de su familia o amigos) y su salud se ven afectados.

¿Qué vida escogería?

Mi vida actual	La vida alternativa	No lo sé
○	○	○

Translation of Figure 5:

Translation of Figure 5:

You indicated that you would prefer the **alternative situation** (with **perfect health and 1050€** per month for consumption) over your **actual situation** (with **your actual health and 1200€** per month for consumption).

In the following table, we present you a another choice, where we have **modified the level of consumption** in the alternative life.

	Your life	**Alternative life**
Consumption	1200€	900€
Health	actual health	perfect health

Consider the two options presented in the table.

You can choose between the **alternative life**, with **perfect health and 900€** per month for your consumption, and **your actual life**. Again, assume that all other aspects of your life remain unchanged, only your personal consumption (not that of your family or friends) and your health are affected.

Which life would you choose?

My actual life	**The alternative life**	**I don't know**
○	○	○

Figure 6 OE question, example for $c_i = 1200$

Ahora, nos gustaría saber cuál es el **mínimo nivel de consumo** al cual escogería la **vida alternativa**.

Imagine que debe **optar entre vivir una vida alternativa** durante el próximo año, con absolutamente **ningún problema ni complicación de salud y X€ por mes** para su consumo, **o su vida actual** (recuerde no tomar en cuenta el alquiler/hipoteca en el consumo). Asuma que todos los otros aspectos de su vida permanecen inalterados, únicamente su consumo personal (no el de su familia o amigos) y su salud se ven afectados.

	Su Vida	Vida Alternativa
Consumo	1200€	X€
Salud	salud actual	salud perfecta

Considere las dos opciones presentadas en la tabla.

¿Cuál sería el **mínimo valor de consumo** (el mínimo valor de X) al cual **preferiría la vida alternativa?**

Indique la cantidad (en Euros):

Translation of Figure 6:

Translation of Figure 6:

Now, we would like to know which is the **minimum level of consumption** at which you would choose the **alternative life**.

Imagine you had to **choose between living an alternative life** for the next year, in which you have **no health problems or complications at all and X€ per month** for consumption, **or your actual life** (remember not to take into account rent/mortgage payments for your consumption). Assume that all other aspects of your life remain unchanged, only your personal consumption (not that of your family or friends) and your health are affected.

	Your life	**Alternative life**
Consumption	1200€	X€
Health	actual health	perfect health

Consider the two options presented in the table.

Which would be the **minimum value of consumption** (the minimum value of X) at which you would **prefer the alternative life**?

Indicate the quantity (in Euros):

A.2 Data set

Figure 7 Flowchart of filtering process

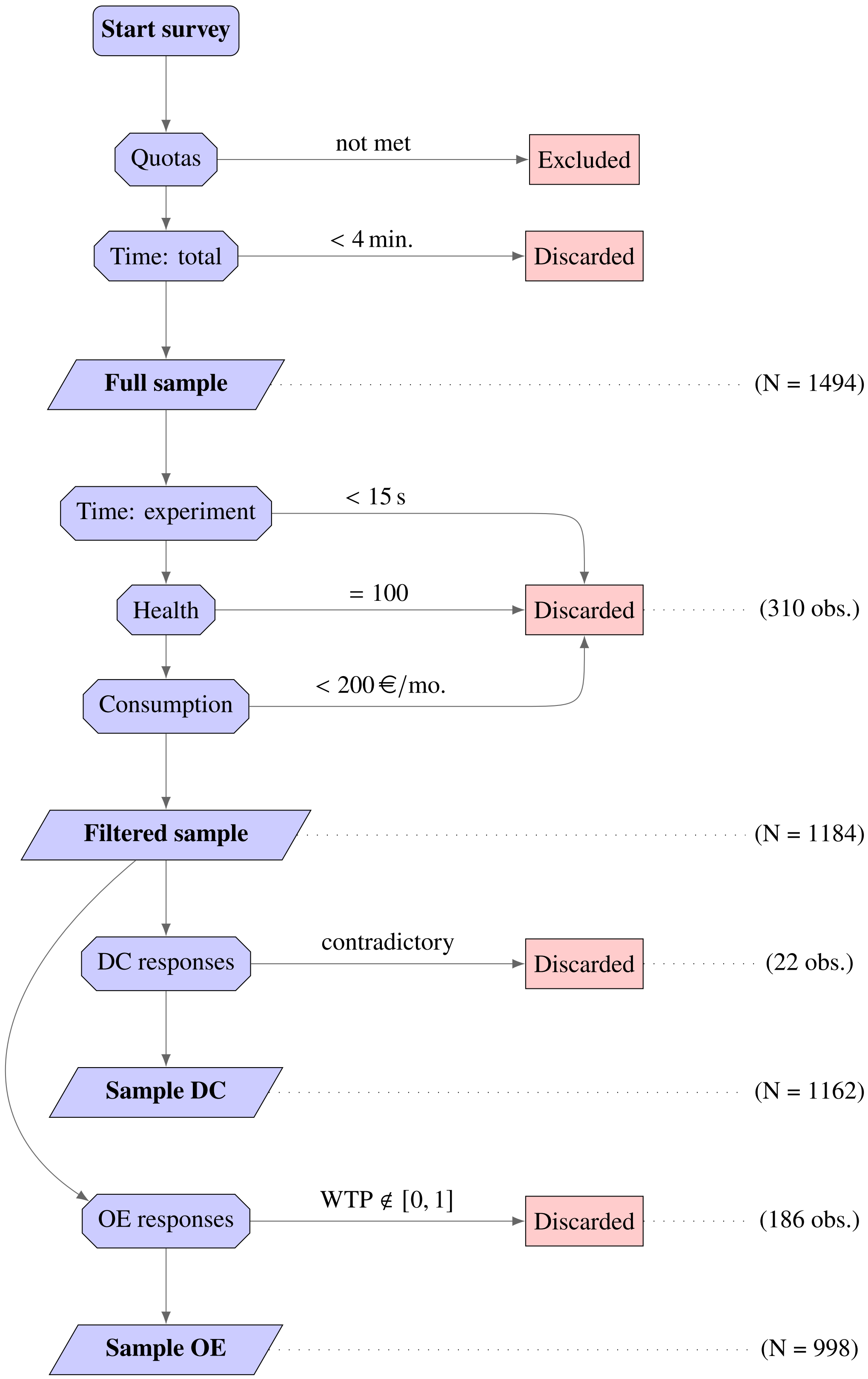


Table 10 Summary statistics of selected variables by treatment

Treatment	Full Sample	Low	Medium	High
Age	44.17	44.80	43.69	44.03
	(10.93)	(10.88)	(10.71)	(11.21)
Female	0.5981	0.5750	0.6026	0.6173
Partnership	0.6073	0.6050	0.6205	0.5964
Household size	3.023	3.030	3.020	3.018
	(1.420)	(1.775)	(1.176)	(1.226)
University education	0.3378	0.3250	0.3410	0.3477
Foreign nationality	0.1345	0.1500	0.1131	0.1399
Political orientation	4.959	5.028	4.741*	5.099
(0–10)	(2.478)	(2.570)	(2.466)	(2.382)
Consumption	950.53	916.37	1048.77*	887.97°
	(1013.51)	(892.75)	(1348.92)	(683.48)
Health	68.97	68.90	69.99	68.03
	(21.46)	(21.46)	(22.14)	(20.79)
Life satisfaction	6.731	6.735	6.841	6.619
	(1.937)	(1.896)	(1.919)	(1.994)
Financial awareness	1.805	1.813	1.777	1.825
(5–1)	(0.970)	(0.977)	(0.969)	(0.964)
Certainty with responses	2.279	2.270	2.221	2.345°
(5–1)	(0.905)	(0.886)	(0.928)	(0.898)
Time: experiment	75.72	75.58	77.69	73.94
	(62.43)	(55.65)	(74.28)	(55.86)
Time: before experiment	399.35	395.97	407.16	395.06
	(275.50)	(282.94)	(289.67)	(253.08)
Total	1184	400	390	394

$^{***}p < 0.01$; $^{**}p < 0.025$; $^{*}p < 0.05$; $^{\circ}p < 0.1$. Variable means, standard deviations in parentheses.

Table 11 Moments and quantiles of OE responses by treatment group

Treatment	Low	Medium	High
Mean	0.2579	0.3189	0.4186
Std. dev.	0.2098	0.2253	0.2686
$q_{0.2}$	0.0889	0.1246	0.1473
$q_{0.5}$	0.2183	0.2929	0.4000
$q_{0.8}$	0.4000	0.5000	0.6725
wtp = 0	11.04%	9.59%	7.62%

A.3 Parametric Analysis: Complete Tables

Table 12 Non-linear regression of open-ended wtp on bids

Dependent variable	Model 6 wtp	Model 7 wtp	Model 8 wtp
Intercept	1.7348***	0.6518**	0.8487***
	(0.3151)	(0.2707)	(0.2201)
Health	−0.0020**	−0.0025**	−0.0021***
	(0.0007)	(0.0008)	(0.0005)
Consumption	−0.0920***	−0.0406°	−0.0439**
(in logs)	(0.0261)	(0.0220)	(0.0182)
b_1	0.2450**	0.4211***	0.2963***
	(0.0826)	(0.0990)	(0.0498)
b_2	0.3329***	0.0290	—
	(0.0422)	(0.1676)	
University	−0.0058	0.0069	0.0013
	(0.0283)	(0.0271)	(0.0224)
Married	−0.0281	−0.0469°	−0.0318
	(0.0302)	(0.0283)	(0.0232)
Household size	0.0148	0.0167*	0.0125°
	(0.0094)	(0.0079)	(0.0064)
Age	−0.0307**	0.0063	−0.0043
	(0.0113)	(0.0103)	(0.0085)
Age^2	0.0003**	−0.0001	0.0000
	(0.0001)	(0.0001)	(0.0001)
Familiar	0.0039	−0.0027	−0.0029
	(0.0106)	(0.0093)	(0.0183)
Evaluation:	−0.0137	0.0002	0.0456
not clear	(0.0227)	(0.0183)	(0.0367)
Evaluation:	0.0065	0.0131	0.0402°
biased	(0.0130)	(0.0114)	(0.0230)
Consent screen:	0.0030	0.0087	0.0385**
seconds	(0.0089)	(0.0078)	(0.0157)
Very certain	−0.0001	−0.0001°	−0.0001*
	(0.0000)	(0.0000)	(0.0001)
Corrections	−0.0544***	−0.0502***	−0.0926***
	(0.0116)	(0.0096)	(0.0194)
J-test statistics	—	7.4341	—
J-test p-value	—	0.9167	—
Sample size	982	982	982
Estimation method	Iter. GMM	Iter. GMM (IV)	Iter. GMM

$^{***}p < 0.001$; $^{**}p < 0.025$; $^{*}p < 0.05$; $^{\circ}p < 0.1$. Std. errors in parentheses, calculated robust to heteroscedasticity.

A.4 Instrument Validity and Robustness Checks

Table 13 Regression of wtp and b_2 on instruments Z_2

Dependent variable	Model (i) wtp	Model (ii) b_2	Model (iii) wtp	Model (iv) b_2
Laptop	0.0023	0.0403	−0.00239	0.0298
	(0.0260)	(0.0246)	(0.02509)	(0.0194)
Locus of control:				
destiny—“not at all”	0.0336	0.0209	0.0382	0.0311
	(0.0323)	(0.0305)	(0.0312)	(0.0242)
own boss—“completely”	−0.0294	0.0110	−0.0326	0.0039
	(0.0269)	(0.0255)	(0.0260)	(0.0202)
hard work—“completely”	−0.0272	0.0262	−0.0310	0.0175
	(0.0215)	(0.0204)	(0.0208)	(0.0161)
internal	0.0138	−0.0395*	0.0127	−0.0420**
	(0.0188)	(0.0178)	(0.0181)	(0.0141)
DC_2 screen:				
clicks	0.00629	0.00239	0.00742	0.00493
	(0.00581)	(0.00550)	(0.00562)	(0.00435)
clicks^2	−0.000156	−0.000169	−0.000170	−0.000201°
	(0.000160)	(0.000152)	(0.000155)	(0.000120)
seconds	0.00000	0.00089	0.00006	0.00103
	(0.00105)	(0.00100)	(0.00102)	(0.00079)
seconds^2	−0.000000	−0.000003	−0.000001	−0.000003
	(0.000006)	(0.000005)	(0.000005)	(0.000004)
DC_1 screen:				
clicks	−0.00827°	−0.00085	−0.00837°	−0.00108
	(0.00454)	(0.00430)	(0.00439)	(0.00340)
clicks^2	0.000252°	0.000150	0.000241°	0.000124
	(0.000138)	(0.000131)	(0.000133)	(0.000103)
seconds	0.000165	−0.000149	0.000198	−0.000076
	(0.000563)	(0.000533)	(0.000544)	(0.000422)
seconds^2	−0.000001	−0.000000	−0.000001	−0.000001
	(0.000001)	(0.000001)	(0.000001)	(0.000001)
Treatment[3]	—	—	−0.00274***	−0.00613***
			(0.00033)	(0.00025)
R^2	0.009	0.017	0.075	0.385
Adj. R^2	-0.005	0.003	0.061	0.376
Sample size	998	998	998	998

***$p < 0.001$; **$p < 0.025$; *$p < 0.05$; °$p < 0.1$. Std. errors in parentheses, calculated robust to heteroscedasticity.

Table 14 Logistic regression on responses to first DC questions

Dependent variable	DC_1 dF/dx	
	Estimate	(Std. error)
Low treatment	0.132***	(0.037)
Medium treatment	0.083*	(0.038)
Health	−0.005***	(0.001)
Consumption (in logs)	−0.035	(0.026)
Female	0.103**	(0.034)
Familiar	−0.108**	(0.041)
Corrections	−0.294***	(0.045)
Laptop	0.103*	(0.052)
Locus of control:		
destiny—“not at all”	0.063	(0.065)
others—“completely”	−0.167°	(0.101)
own boss—“completely”	0.011	(0.057)
hard work—“completely”	0.081*	(0.041)
internal	−0.093**	(0.038)
Consent screen:		
seconds	−0.001*	(0.0004)
DC_1 screen:		
clicks	0.0002	(0.014)
clicks^2	0.001	(0.001)
seconds	−0.001	(0.001)
seconds^2	−0.00000	(0.00000)
DC_2 screen:		
clicks	0.012	(0.013)
clicks^2	−0.0004	(0.0003)
seconds	0.003	(0.002)
seconds^2	−0.00000	(0.00001)
Sample size	982	
Estimation method	Logistic regression	

$^{***}p < 0.001$; $^{**}p < 0.025$; $^{*}p < 0.05$; $^{\circ}p < 0.1$. Std. errors in parentheses, calculated robust to heteroscedasticity.

A.5 Residual terms in convex model

In this section, we provide a detailed explanation of the residual terms from the convex model employed in Section 4.2.2. Recall the expression from Equation 5:

$$\text{wtp}_i = (1 - \gamma - \omega + \gamma\,\omega)\,\beta_X\,X_i + \gamma\,(1 - \omega)\,b_{1i} + \omega\,b_{2i} + \eta_i^*.$$

This equation combines three terms:

$$\begin{aligned}
\text{wtp}_{0,i} &= \beta_X\,X_i + \varepsilon_{0,i} \\
\text{wtp}_{1,i} &= (1 - \gamma)\,\text{wtp}_{0,i} + \gamma\,b_{1,i} + \eta_{1,i} \\
\text{wtp}_{2,i} &= (1 - \omega)\,\text{wtp}_{1,i} + \omega\,b_{2,i} + \eta_{2,i}.
\end{aligned}$$

We assume that $\varepsilon_{0,i}$ is a random variable that follows a normal distribution with zero mean. After observing b_1 individuals update their wtp to $\text{wtp}_{1,i}$. The residual term $\eta_{1,i}$ comprises two components: a deterministic part, which depends on survey-related features captured in Q_i, and an independent random error term $\varepsilon_{1,i}$. Formally, $\eta_{1,i} = f(\delta', Q_{1,i}) + \varepsilon_{1,i}$, where δ' is a vector of unknown parameters, and $f(.)$ is an unknown function. The first term on the right-hand side represents deterministic factors that influence how individuals update their $\text{wtp}_{1,i}$ but do not explain the prior $\text{wtp}_{0,i}$. While it is not the focus of our model, we want to control for these components to reduce noise. Similarly, when individuals respond to the second DC, there is a deterministic and random component of the residual term. Hence, we have: $\eta_{2,i} = f(\delta'', Q_{2,i}) + \varepsilon_{2,i}$.

Assuming that $f(.)$ is a linear function and that $Q_{1,i} = Q_{2,i}$—the same deterministic components influence the adjusted wtp after seeing the first and second DC questions, although not necessarily by the same magnitude, i.e. $\delta' \neq \delta''$—and since we are not interested in differentiating the magnitude of these effects separately—the aim is to reduce noise—we can combine both terms to obtain $\eta_i^* = (2-\omega)\,\delta\,Q_i + \varepsilon_i^*$, where: $\varepsilon_i^* = (1-\omega)\,[(1 - \gamma)\,\varepsilon_{0,i} + \varepsilon_{1,i}] + \varepsilon_{2,i}$. Since all the terms ($\varepsilon_{0,i}$, $\varepsilon_{1,i}$ and $\varepsilon_{2,i}$), follow normal distributions with zero mean, ε_i^* also follows a normal distribution with zero mean, provided that $\varepsilon_{0,i}$, $\varepsilon_{1,i}$ and $\varepsilon_{2,i}$ are independent. The final model to be estimated is:

$$\text{wtp}_i = (1 - \gamma - \omega + \gamma\,\omega)\,\beta_X\,X_i + \gamma\,(1 - \omega)\,b_{1i} + \omega_i\,b_{2i} + (2 - \omega)\,\delta\,Q_i + \varepsilon_i^*.$$

Variables that may explain $\text{wtp}_{0,i}$, such as health or consumption, should be included in X_i. Survey-related features, such as response certainty or time spent answering, can be controlled for by including them in Q_i. The variables in Q_i are expected to align with those specified in Equation 1 and Equation 2. It is important to note that omitting the deterministic component of the residual terms η results in a more parsimonious model. While the estimates remain consistent and qualitatively similar omitting this term, the precision of the estimation decreases, with larger standard errors. This robustness analysis is presented in Section A.6 of the Appendix (Table 18).

A.6 Robustness Analysis

Tables 15, 16, and 17 present a sensitivity analysis of the primary models from Sections 4.1 (Models 1 and 2), 4.2.1 (Models 3, 4, and 5), and 4.2.2 (Models 6, 7, and 8) respectively, by varying the definition of the dependent variable. The original models are also included for comparison. Models labelled with a letter "a" use "wtp (uncorrected)" as the dependent variable, keeping the initial value provided in the OE question without any subsequent corrections made by respondents. Hence, Model 1 in Table 15 is the exact same as Model 1 from Table 6, while Model 1a replicates Model 1 using "wtp (uncorrected)" as the dependent variable. Models labelled with a letter "b" use the variable "wtp (no errors)" as the dependent variable, which adjusts the value to compute wtp only for individuals who corrected their response due to inattention or misunderstanding while retaining the original first value for those who indicated they changed their mind (for these cases, the first value is kept). In Table 18 an additional sensitivity analysis is performed where Models 6, 7, and 8 are replicated assuming no deterministic component of the residual term η, labelled with the names Models 6c, 7c, and 8c respectively.

Table 15 Robustness check—regression of wtp on treatments

wtp	Model 1 corrected	Model 1a uncorrected	Model 1b no errors	Model 2 corrected	Model 2a uncorrected	Model 2b no errors
Intercept	0.4803**	0.5147**	0.4803**	0.4459**	0.4752**	0.4460**
	(0.1560)	(0.1570)	(0.1580)	(0.1551)	(0.1571)	(0.1580)
Health	−0.0015***	−0.0015***	−0.0014***	−0.0015***	−0.0014***	−0.0014***
	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
Consumption (in logs)	−0.0224°	−0.0244°	−0.0198	−0.0222°	−0.0247°	−0.0199
	(0.0128)	(0.0129)	(0.0130)	(0.0125)	(0.0128)	(0.0128)
Low treatment	−0.0662***	−0.0708***	−0.0670***	−0.0525**	−0.0472°	−0.0466*
	(0.0168)	(0.0174)	(0.0169)	(0.0230)	(0.0242)	(0.0234)
× university	—	—	—	−0.0530	−0.0640°	−0.0564
				(0.0343)	(0.0360)	(0.0350)
× familiar	—	—	—	0.0329	0.0015	0.0060
				(0.0426)	(0.0444)	(0.0425)
High treatment	0.0987***	0.0803***	0.0920***	0.1605***	0.1390***	0.1538***
	(0.0190)	(0.0194)	(0.0193)	(0.0260)	(0.0263)	(0.0263)
× university	—	—	—	−0.1186**	−0.1176**	−0.1226**
				(0.0406)	(0.0417)	(0.0413)
× familiar	—	—	—	−0.0906*	−0.0804°	−0.0870°
				(0.0455)	(0.0469)	(0.0463)
University	0.0001	0.0038	0.0018	0.0547*	0.0633**	0.0594*
	(0.0159)	(0.0164)	(0.0161)	(0.0260)	(0.0273)	(0.0265)
Familiar	−0.0068	0.0034	0.0039	0.0125	0.0305	0.0303
	(0.0187)	(0.0192)	(0.0188)	(0.0288)	(0.0302)	(0.0294)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.132	0.117	0.121	0.148	0.130	0.136
Adj. R^2	0.119	0.103	0.107	0.131	0.112	0.118
Sample size	982	922	956	982	922	956
Est. method	OLS	OLS	OLS	OLS	OLS	OLS

$^{***}p < 0.001$; $^{**}p < 0.025$; $^{*}p < 0.05$; $^{\circ}p < 0.1$. Std. errors in parentheses, calculated robust to heteroscedasticity.

Table 16 Robustness check—linear regression of wtp on bids

wtp	Model 3 corrected	Model 3a uncorrected	Model 3b no errors	Model 4 corrected	Model 4a uncorrected	Model 4b no errors	Model 5 corrected	Model 5a uncorrected	Model 5b no errors
Intercept	0.2688°	0.3277^{*}	0.2840°	0.3400^{*}	0.3687^{**}	0.3375^{*}	0.3591^{**}	0.3946^{**}	0.3592^{**}
	(0.1459)	(0.1467)	(0.1480)	(0.1553)	(0.1518)	(0.1545)	(0.1545)	(0.1555)	(0.1563)
Health	-0.0006°	-0.0006°	-0.0006°	-0.0013^{**}	-0.0011^{**}	-0.0012^{**}	-0.0015^{***}	-0.0014^{***}	-0.0014^{***}
	(0.0003)	(0.0004)	(0.0003)	(0.0005)	(0.0005)	(0.0005)	(0.0004)	(0.0004)	(0.0004)
Consumption (in logs)	-0.0239^{*}	-0.0278^{**}	-0.0222°	-0.0224°	-0.0256^{*}	−0.0203	-0.0221°	-0.0242°	−0.0195
	(0.0121)	(0.0122)	(0.0124)	(0.0124)	(0.0125)	(0.0127)	(0.0126)	(0.0128)	(0.0128)
b_1	0.0101	−0.0327	−0.0108	0.3434°	0.2261	0.2930	0.4328^{***}	0.3902^{***}	0.4157^{***}
	(0.0593)	(0.0587)	(0.0602)	(0.1821)	(0.1743)	(0.1805)	(0.0499)	(0.0507)	(0.0503)
b_2	0.4673^{***}	0.4656^{***}	0.4735^{***}	0.0988	0.1806	0.1362	—	—	—
	(0.0414)	(0.0417)	(0.0418)	(0.1919)	(0.1837)	(0.1904)			
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.132	0.116	0.121	0.175	0.190	0.180	0.247	0.234	0.241
Adj. R^2	0.119	0.102	0.108	0.163	0.176	0.167	0.236	0.222	0.229
Sample size	982	922	956	982	922	956	982	922	956
Est. method	OLS	OLS	OLS	IV	IV	IV	OLS	OLS	OLS

$^{***}p < 0.001$; $^{**}p < 0.025$; $^{*}p < 0.05$; $^{\circ}p < 0.1$. Std. errors in parentheses, calculated robust to heteroscedasticity.

Table 17 Robustness check—non-linear regression of wtp on bids

wtp	Model 6 corrected	Model 6a uncorrected	Model 6b no errors	Model 7 corrected	Model 7a uncorrected	Model 7b no errors	Model 8 corrected	Model 8a uncorrected	Model 8b no errors
Intercept	1.7348***	0.3072	0.2971	0.6518**	0.6516**	0.6320**	0.8487***	0.2478	0.2851
	(0.3151)	(0.2597)	(0.2840)	(0.2707)	(0.2498)	(0.2630)	(0.2201)	(0.2174)	(0.3122)
Health	−0.0020**	−0.0024***	−0.0024***	−0.0025**	−0.0022**	−0.0023**	−0.0021***	−0.0020***	−0.0023**
	(0.0007)	(0.0006)	(0.0007)	(0.0008)	(0.0007)	(0.0008)	(0.0005)	(0.0005)	(0.0007)
Consumption (in logs)	−0.0920***	−0.0555**	−0.1695***	−0.0406°	−0.0396°	−0.0362°	−0.0439**	0.0183	0.1530***
	(0.0261)	(0.0222)	(0.0292)	(0.0220)	(0.0208)	(0.0217)	(0.0182)	(0.0179)	(0.0292)
b_1	0.2450**	0.3495***	0.3922***	0.4211***	0.3465**	0.3863***	0.2963***	0.2904***	0.2612***
	(0.0826)	(0.0625)	(0.0616)	(0.0990)	(0.1139)	(0.1074)	(0.0498)	(0.0511)	(0.0744)
b_2	0.3329***	0.1022**	−0.0695	0.0290	0.0824	0.0604	—	—	—
	(0.0422)	(0.0444)	(0.0516)	(0.1676)	(0.1625)	(0.1675)			
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
IV Approach	No	No	No	Yes	Yes	Yes	No	No	No
J-test statistics				7.4341	9.2058	9.5056			
J-test p-value				0.9167	0.8176	0.7974			
Sample size	982	922	956	982	922	956	982	922	956
Est. method	It. GMM	It. GMM	It. GMM	It. GMM	It. GMM	It. GMM	It. GMM	It. GMM	It. GMM

$^{***}p < 0.001$; $^{**}p < 0.025$; $^{*}p < 0.05$; $^{\circ}p < 0.1$. Std. errors in parentheses, calculated robust to heteroscedasticity.

Table 18 Robustness check—non-linear regression of wtp on bids (no η)

Dep. variable	Model 6 wtp	Model 6c wtp	Model 7 wtp	Model 7c wtp	Model 8 wtp	Model 8c wtp
Intercept	1.7348^{***}	3.7744^{***}	0.6518^{**}	0.5777^{*}	0.8487^{***}	1.1576^{***}
	(0.3151)	(0.3676)	(0.2707)	(0.2633)	(0.2201)	(0.2221)
Health	-0.0020^{**}	-0.0023^{***}	-0.0025^{**}	-0.0023^{**}	-0.0021^{***}	-0.0022^{***}
	(0.0007)	(0.0006)	(0.0008)	(0.0008)	(0.0005)	(0.0005)
Consumption (in logs)	-0.0920^{***}	-0.1639^{***}	-0.0406°	-0.0460^{*}	-0.0439^{**}	-0.0585^{**}
	(0.0261)	(0.0249)	(0.0220)	(0.0216)	(0.0182)	(0.0182)
b_1	0.2450^{**}	0.1716^{*}	0.4211^{***}	0.3593^{**}	0.2963^{***}	0.2886^{***}
	(0.0826)	(0.0779)	(0.0990)	(0.1210)	(0.0498)	(0.0501)
b_2	0.3329^{***}	0.1781^{***}	0.0290	0.1235	—	—
	(0.0422)	(0.0466)	(0.1676)	(0.1734)		
Controls	Yes	Yes	Yes	Yes	Yes	Yes
IV Approach	No	No	Yes	Yes	No	No
Det. residual (η)	Yes	No	Yes	No	Yes	No
J-test statistics			7.4341	9.1644		
J-test p-value			0.9167	0.8204		
Sample size	982	988	982	988	982	988
Est. method	It. GMM	It. GMM	It. GMM)	It. GMM	It. GMM	It. GMM

$^{***}p < 0.001$; $^{**}p < 0.025$; $^{*}p < 0.05$; $^{\circ}p < 0.1$. Std. errors in parentheses, calculated robust to heteroscedasticity.

A.7 Equivalent consumption corrections

Figure 8 Cumulative distribution functions of stated WTP by treatment

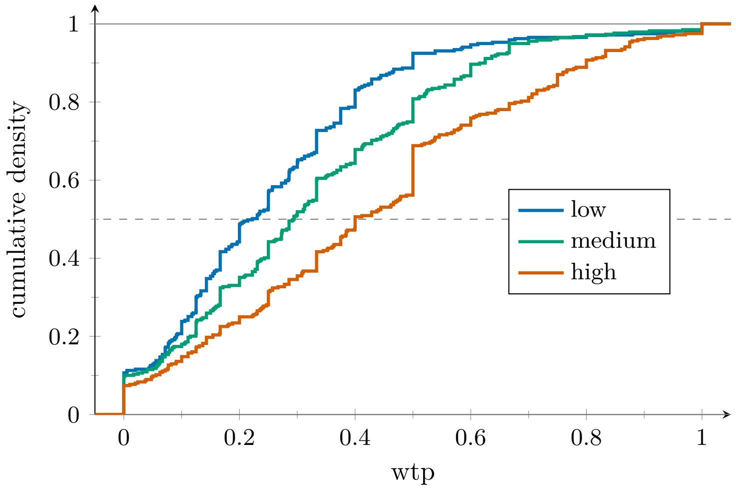


Figure 9 Cumulative distribution functions of predicted WTP by treatment