Behavioral drivers or economic incentives? Toward a better understanding of elicitation effects in stated preference studies

Christian A. Vossler*
Department of Economics and Howard H. Baker Jr. Center for Public Policy,
University of Tennessee, Knoxville, TN, USA
cvossler@utk.edu

and

Ewa Zawojska
Faculty of Economic Sciences
University of Warsaw, Warsaw, Poland
ewa.zawojska@uw.edu.pl

Accepted for publication by the *Journal of the Association of Environmental and Resource Economists* on 10/01/2019. https://www.journals.uchicago.edu/loi/jaere

*Correspondence should be directed to Christian Vossler, Department of Economics, 527F Stokely Management Center, The University of Tennessee, Knoxville, TN 37996-0550. E-mail: cvossler@utk.edu. Telephone: 865-974-1699. Fax: 865-974-4601. We thank Adrienne Sudbury for her research assistance. Robert Johnston and two anonymous reviewers provided valuable comments. This research was undertaken while Ewa Zawojska was a visiting PhD student at the University of Tennessee, under the generous support of the National Science Centre in Poland (Opus 13, grant no. 2017/25/B/HS4/01076; Etiuda 4, grant no. 2016/20/T/HS4/00013). This research has been supported by a grant from the U.S. Environmental Protection Agency's Science to Achieve Results (STAR) program.

Behavioral drivers or economic incentives? Toward a better understanding of elicitation effects in stated preference studies

Abstract: Survey-based welfare measures for public goods are often very sensitive to value elicitation methods, e.g., whether the elicitation is framed as an up-or-down vote or an open-ended willingness-to-pay (WTP) question. Leading hypotheses for elicitation effects are tied to either behavioral factors (e.g., psychological cues) or differences in perceived economic incentives. Past research that emphasizes behavioral drivers does not precisely control economic incentives, potentially confounding inferences. We conduct an experiment that controls economic incentives through incentive-compatible elicitation methods, and compare single binary choice, double-bounded binary choice, payment card, and open-ended response formats. Our experiment retains important field survey characteristics, including the valuation of an environmental public good with passive-use value. All formats elicit statistically indistinguishable WTP distributions, suggesting that behavioral factors may not be the primary drivers of elicitation effects. To the extent that our laboratory methods can be mirrored in the field, this offers a pathway for mitigating elicitation effects.

Keywords: contingent valuation; mechanism design; experiment; voting; elicitation effects; convergent validity

JEL classification: Q51; C92; D82; H41

1. Introduction

In the context of using stated preference surveys to estimate monetary values for non-market goods, such as a change in air quality or a land conservation project, researchers are charged with the challenging task of providing valid welfare measures to inform decision-making processes. Although many threats to the validity of stated preference measures exist, as discussed in Johnston et al. (2017), much research centers on value elicitation methods, predominantly the question response format (e.g., a binary choice or an open-ended willingness-to-pay (WTP) question). The stylized fact in the literature is that estimated values are very sensitive to the question format, giving rise to so-called "elicitation effects" (e.g., Bateman et al., 2001; Cameron et al., 2002; Champ and Bishop, 2006).[1] Critics interpret this as evidence that stated preference surveys lack external validity (e.g., McFadden and Leonard, 1993). Myriad elicitation formats continue to be used in practice, and this raises concerns for both academics and policymakers (see, for example, Kling et al., 2012; Bishop et al., 2017). As a step forward in understanding the underlying drivers of elicitation effects, in this study we design and implement a novel experiment involving an environmental public good that allows us to compare a set of popular question formats while holding fixed economic incentives. In doing so, we investigate whether hypothesized behavioral factors are of first-order importance in explaining elicitation effects. Further, mechanisms similar to the ones we implement have been shown to reveal demand in induced-value experimental settings. As such, our investigation provides a more stringent test of these mechanisms in a field-resembling environment, while providing insight on survey design.

[1] Other methodological choices related to value elicitation have been shown to (sometimes) influence valuations, such as the payment vehicle, the description of the commodity, and the use of "cheap talk" scripts (e.g., Boyle, 1989; Champ et al., 2002; Ami et al., 2011).

Leading explanations for elicitation effects in stated preference surveys are tied to either behavioral arguments or differences in perceived economic incentives across formats (see Bateman et al., 2001).[2] Focusing first on behavioral arguments, responses to single binary choice questions and other posted-price formats might be subject to "anchoring" or "reference point" bias, which can arise, for instance, from a respondent's unfamiliarity with the elicitation method or with the good being valued. This has been suggested as an explanation for observed differences between single binary choice and open-ended responses (e.g., Green et al., 1998; O'Conor et al., 1999) and for differences between single and double-bounded binary choice responses (e.g., Whitehead, 2002; Scheufele and Bennett, 2013), among others. Hanemann (1995) argues that bids (prices) could be construed as quality signals. This can lead to differences between closed- and open-ended elicitation mechanisms, as for the latter there are no posted prices. Welsh and Poe (1998) show that elicitation effects may be driven by response (un)certainty, with respondents to single binary choice questions being less certain about their choices relative to respondents to payment card and open-ended questions. Similar links between elicitation effects and response certainty have been documented in other work, e.g., Ready et al. (2001). Frew et al. (2003) speculate that observed differences in single binary choice valuations relative to open-ended and payment card formats may be in part attributable to "yea-saying."

Turning now to economic incentives, as carefully argued by Carson and Groves (2007), alternatives to a single binary choice question are commonly believed to incentivize respondents to misrepresent their true preferences.[3] For instance, a respondent to an open-ended question may perceive she can influence the cost to her upon project implementation, thus incenting the under-

[2] Although it has received less attention in the literature, another explanation is that differences arise due to the misspecification of econometric models (see, for example, Huang and Smith, 1998).

[3] Of course, arguments based on economic incentives implicitly assume that survey respondents perceive there to be potential consequences tied to their choices (in other words, the survey is assumed to be consequential).

revelation of demand. On the other hand, if she believes that cost is fixed, but that the chance a considered project is implemented is increasing in total (or average) WTP, she will overstate her WTP as long as it is higher than this fixed cost, and will otherwise state a WTP of zero. In the case of the double-bounded binary choice format, facing the second question may trigger a belief that cost is uncertain (but exogenous to the respondent's decisions) or, similar to the open-ended setting, a belief that the respondent can influence cost. If she holds the former belief, she may then respond based on her expectation of the true cost (e.g., the average of the first and second costs presented), rather than the stated cost. When presented a sequence of questions involving related goods, or different prices for the same good (e.g., as in a payment card), this likewise can give rise to beliefs that lead to a loss of incentive compatibility (Vossler et al., 2012).

When interpreting results from past studies, it is difficult to determine whether observed elicitation effects are due to behavioral factors or differences in (perceived) economic incentives across question formats. That is, evidence on elicitation effects in the public goods valuation setting comes from stated preference surveys, and from actual payment experiments involving donation mechanisms. In these settings it is not possible to precisely control economic incentives when comparing formats, and therefore differences in economic incentives serve to potentially confound tests of hypotheses based on behavioral theories. In this study, in order to better understand the drivers of elicitation effects, we design an innovative experiment that controls economic incentives through the use of incentive-compatible elicitation methods. Specifically, in a setting where homegrown values for an environmental public good are elicited, we test whether behavioral drivers are important in a comparison of four question formats – single binary choice, double-bounded binary choice, payment card, and open-ended.

The broader experimental literature provides no clear guidance on whether we should expect elicitation effects to persist after controlling economic incentives. Poe (2016) gives many examples of "behavioral anomalies" found in both the revealed preference and stated preference literatures. There are also cases where alternative incentive-compatible mechanisms are not empirically equivalent. For example, comparisons between the Becker-DeGroot-Marschak (BDM; Becker et al., 1964) mechanism and the second-price auction generally reveal differences (e.g., Rutström, 1998). As one possible explanation, only the latter mechanism promotes competition between players, leading to behavioral motivations such as a "joy of winning" and "spite" (Cooper and Fang, 2008). On a related note, induced-value experiments suggest some incentive-compatible mechanisms are not demand revealing.[4] As a source of contrast to the above findings, accumulated evidence demonstrates that behavioral phenomena such as anchoring and endowment effects may become less pronounced or even eliminated when one uses careful experimental procedures, including incentive-compatible elicitation mechanisms (e.g., Plott and Zeiler, 2005; Fundenberg et al., 2012; Alevy et al., 2015).

The results of our investigation are of potential importance to both researchers and policy makers. If our experimental design does not eliminate elicitation effects, this emphasizes the importance of behavioral factors. Importantly, this suggests that elicitation effects arise even in a setting with direct financial consequences and, accordingly, the accumulated evidence from stated preference surveys does not directly imply that elicited values are biased. On the other hand, if we do not find elicitation effects, this implies that elicitation effects in the field are largely driven by differences in perceived economic incentives. Moreover, to the extent our incentive-compatible

[4] Examples include a "pivot" version of the Vickery-Clarke-Groves mechanism (Attiyeh et al., 2000), and a payment card format commonly used in the stated preference literature (Vossler and McKee, 2006). Presumably if we included such mechanisms in comparisons involving homegrown values, they would likewise lead to empirical differences.

elicitation methods can be used to inform the design of field surveys, this result suggests the possibility of mitigating elicitation effects in the field.

Much of the literature on elicitation effects focuses on the four formats investigated here, and behavioral theories have been proposed to explain observed differences in elicited values between them. Moreover, we compare formats that have been shown to be demand revealing in experiments involving induced values, which minimizes the chance that any elicitation effects we find might instead be attributable to a misunderstanding of economic incentives.[5] It is important to note that the hypothesized behavioral drivers are tied to the processes through which people form valuations for a non-market good, and so are effectively neutralized in induced-value settings.

We implement elicitation formats so that truthful preference revelation constitutes the single best response strategy; that is, all formats are incentive compatible. In the case of the double-bounded binary choice, payment card and open-ended formats, this is achieved by introducing uncertainty over the cost an individual has to pay in the event the project is funded. The resulting mechanisms can be viewed as repeated binary choice mechanisms with a random cost selection rule. Azrieli et al. (2018) show that the incentive compatibility of such mechanisms arises under a monotonicity assumption. We hold constant ancillary characteristics such as framing and the payment vehicle to rule out other potential confounds.

Important for the generalizability of our lab results to the field, the valuation task shares characteristics of many stated preference surveys. Specifically, our experiment elicits valuations for an actual environmental public good. As such, unlike in an induced-value setting, our design does not mitigate the effects of possible behavioral drivers such as anchoring or heterogeneous

[5] Taylor et al. (2001), Vossler and McKee (2006), and Collins and Vossler (2009) provide evidence on demand revelation for the single binary choice format. Evidence on the double-bounded format comes from Carson et al. (2009). Evidence on the payment card and open-ended formats can be found in Vossler and McKee (2006) and Messer et al. (2010), respectively.

interactions between response uncertainty and elicitation formats. Our study involves forestation of agricultural land in a distant location; as a result, the elicitation mainly captures passive-use values. Of course, the prominent conceptual advantage of stated preference surveys over observational studies is the ability to measure welfare changes associated with passive use. Participants in our study are presumed to be unfamiliar with the good, particularly, with the task of evaluating such a good. Further, mirroring the field, whether individual costs differ across participants is left ambiguous to participants.

Our main result is that we find no statistical evidence of elicitation effects. Estimates of mean WTP as well as empirical WTP distributions are statistically identical across the four elicitation formats. This is strong evidence of convergent validity. One implication of this result is that behavioral factors and biases postulated to give rise to elicitation effects, such as anchoring and complexity, may be of second-order importance. To the extent the random cost selection mechanisms we implemented in the lab can be conceptually paralleled in the field, our results further suggest the possibility that elicitation effects can be mitigated. On this point, we provide a discussion of the lab-to-field generalizability of our results and methods later in the article.

2. Experimental design

2.1 Valuation scenario

The experiment elicits preferences toward funding a tree-planting project. The project involves planting and maintaining 160 trees (about ¼-acre) on agricultural land along the Mississippi River Valley. To achieve this outcome, we collaborated with the organization GreenTrees. The organization currently carries out various tree-planting projects across a long stretch of the Mississippi River Valley. Participants receive information on the broad benefits of tree planting. In addition, based on analyses of other projects undertaken by GreenTrees in the region, we present participants with estimates of increased water storage, avoided nutrient runoff, and CO_2 capture that would occur if the specific project considered in our experiment were funded. The experiment instructions, which describe additional details of the project, are included in the Online Appendix.

The tree-planting project was chosen to enhance the generalizability of the study. As in many stated preference surveys, this project is a non-market, public good. Values for the project should be predominantly tied to passive use, as the Mississippi River Valley is hundreds of miles away from where study participants currently live. Also typical of valuation scenarios in contingent valuation surveys, it is unlikely that our participants have experience funding tree-planting projects. This characteristic is of particular importance in the context of our research in which, as discussed in the introduction, we hypothesize that possible value cues provided by the elicitation format may influence valuations. Further, although individuals can seek out opportunities to fund tree plantings or other carbon offsets, it is improbable that participants envision another opportunity to fund collectively a project of this scope. On their website, GreenTrees does not provide an explicit way for individuals to support tree plantings, which are largely funded by corporations in exchange for carbon offsets.

2.2 Experimental treatments

In designing this test of elicitation effects, aside from inherent differences in elicitation formats, we aim to hold fixed important characteristics of the elicitations as they may influence valuations. In particular, across all formats: (1) value elicitation questions are framed as referenda and rely on a majority-vote implementation rule; (2) it is ambiguous as to whether the individual cost of funding the project, which is displayed to participants privately, varies across participants (similarly to actual field applications, the individual costs in our experiment do vary); and (3) it is common knowledge that neither the total amount paid to GreenTrees (if funded) nor the size of the project (i.e., the number of trees planted) depends on the amount of money collected from participants. The instructions emphasize that the total cost of the project has been pre-negotiated, and that the project will be subsidized by the experimenters if the sum of individual costs collected upon the referendum passing is less than the total cost of the project.[6] While the exact amount of the total project cost is not disclosed, participants are explicitly informed about the number of trees planted. The design helps avoid notions of "fair share" pricing, possible collection of excess funds, and speculations that the scope of the project depends on the amount collected.

The experiment includes four split-sample treatments, each involving a separate elicitation mechanism: single binary choice, double-bounded binary choice, payment card, or open-ended. The wording used to explain the elicitation mechanisms is as similar as possible across treatments, only varying as required to convey specifics of the mechanisms. We now describe how each mechanism is operationalized, and provide theoretical justification.

2.2.1 Single binary choice

The single binary choice (SBC) treatment involves a simple up-or-down vote on whether to fund the project at a specific individual cost. The exact wording of the value elicitation question is, "If passage of the referendum cost you $x, are you in favor of funding the tree planting project?" As standard in stated preference surveys employing this format, the individual cost varies across voters to make identification of WTP possible. A binding binary choice referendum with a

[6] In the experiment, the sum of individual costs collected upon the referendum passing is always lower than the pre-negotiated total cost of implementing the project.

majority-vote implementation rule is well known to be incentive compatible under weak assumptions (Farquharson, 1969).

2.2.2 Double-bounded binary choice

The double-bounded binary choice (DB) treatment extends the SBC elicitation by presenting participants with two separate referenda that differ only in the stated individual cost. That there are two votes is common knowledge, although the two costs are not known in advance. The second referendum is displayed once all votes in the first referendum are submitted. The wording of each value elicitation question is identical to the SBC treatment. To break possible strategic ties between the two votes, one of the two referenda is randomly selected to be binding after all decisions are made. To understand the incentive properties of this mechanism, we draw from the broader theoretical literature that analyzes incentives in experimental games. In particular, attention has been devoted to games involving a sequence of binary choices, one of which is selected at random to be binding. Cox et al. (2015) prove that such mechanisms are incentive compatible for all theories that satisfy the reduction of compound lotteries axiom and the independence axiom. Azrieli et al. (2018) instead demonstrate that incentive compatibility arises under a weaker monotonicity assumption, and that the independence axiom is only needed when the reduction axiom is also assumed.[7]

2.2.3 Payment card

[7] This theory encompasses not only individual-choice mechanisms but also strategic games where payoffs depend on the joint actions of a group of players. This characterizes voting in our experiment. An important requirement is that each game (vote), if analyzed in isolation, represents an incentive compatible elicitation. Therefore, we need for each vote (cost) that may be selected, that a voter perceives her choice can probabilistically determine the outcome.

In the payment card (PC) treatment, participants are provided with several possible individual costs of funding the project, and the elicitation is cast as an up-or-down vote to each cost. This reflects some field implementations of the mechanism (for example, see Bateman et al., 2005), although it is more common to ask respondents to circle, from the provided costs, the highest amount they are willing to pay.[8] Framing the elicitation as a set of (independent) votes to each cost amount enhances transparency of the mechanism, while more closely resembling the other mechanisms we study. Moreover, this form of implementation makes the payment card similar to a binary discrete choice experiment (binary DCE). Conceptually, this generalizes the DB elicitation to more than two referenda. In implementation, and as relayed to voters, one of the referenda is randomly selected to be binding.[9] As in the DB treatment, the random selection process breaks the strategic link across decisions in the sense that participants cannot influence the cost paid. The wording of the referendum is identical to the prior two treatments. Assumptions for incentive compatibility are identical to those for the DB treatment.

2.2.4 Open-ended

The open-ended (OE) format elicits a point estimate of value. The wording of the referendum is revised to ask: "What is the highest amount that you would pay and still vote in favor of funding the tree planting project?" To parallel the other mechanisms, as well as to provide information to enhance truthful demand revelation, the format is described as a concise way to learn the range of possible individual cost amounts for which a person would vote "yes" or "no." It is further emphasized that, as in the PC treatment, the binding individual cost is randomly

[8] Vossler and McKee (2006) test the latter form of a payment card using an induced-value experiment, and conclude that it is not demand revealing.

[9] To parallel the SBC and DB mechanisms, the individual cost of the binding referendum in the PC treatment varies across people, and this is not made explicit in the instructions.

determined after all decisions are submitted. Once the cost is drawn, the stated valuation is compared to the cost. If the valuation is equal to or higher than the cost, this is a "yes" vote in an up-or-down referendum at this cost. Otherwise, this is a "no" vote. Theoretically, our OE mechanism is the Random Price Voting Mechanism (RPVM) developed by Messer et al. (2010).

In implementation, neither the range nor probability distribution of possible individual costs is made explicit. This is a deliberate design choice to reflect field conditions. Further, this helps assure incentive compatibility. As discussed by Azrieli et al. (2018), it does not matter theoretically if the experimenter uses a randomization device that is ambiguous to participants. Similar to the BDM mechanism, incentive compatibility of the RPVM does require that a voter perceives there is a positive probability that the realized cost is equal to her valuation.[10]

Messer et al. (2010) demonstrate incentive compatibility of the RPVM under expected utility. However, as the stated OE valuation theoretically maps into a continuum of "yes" or "no" votes to referenda distinguished by the cost, the mechanism can be analyzed in a similar manner to the DB and PC mechanisms; that is, it can be interpreted as a sequence of binary choices, one of which is chosen at random to be binding. Therefore, the RPVM should be incentive compatible when the assumption of expected utility is relaxed.

2.3 Design parameters

The bid designs (i.e., procedures for selecting individual costs) used in the SBC, DB and PC elicitations are informed by a pre-test and a pilot session (*n* = 46), both involving OE elicitations. These results yield an observed distribution of WTP reasonably approximated by a

[10] To illustrate this, suppose a voter's valuation is \$5 and she believes with certainty that the highest possible individual cost is \$3. In this case, the voter is indifferent between stating any amount greater than or equal to \$3. To help mitigate such beliefs, the instructions explicitly state that the range of possible costs vary from "very low to very high amounts."

normal distribution with a mean of \$3 and a standard deviation of 1.75. Although research on bid designs for SBC and DB elicitations suggests that a small number of bids placed, loosely speaking, sufficiently away from the tails of the distribution is most efficient (Alberini, 1995), such designs assume that the underlying WTP distribution is known. As we anticipated treatment effects, we instead utilize a large set of bids that (roughly) span the 20th to 80th percentiles of the WTP distribution observed in the pre-test and pilot. Specifically, for the SBC treatment, we draw bids randomly from the vector {\$1, \$2, \$3, \$4, \$5, \$6} with equal probabilities. This corresponds with the general rule-of-thumb given in Kanninen (1995) that bids should not fall outside the 15th and 85th percentiles of the distribution.

As standard in the literature, to enhance the efficiency of the DB design, the individual cost in the second referendum depends on whether a participant votes "yes" or "no" in the first referendum. Mechanically, if a voter responds "no" to the first cost, the second vote involves a cost randomly drawn from a set of lower cost amounts available (e.g., if a voter says "no" to \$3 in the first referendum, a cost of \$1 or \$2 is drawn with equal probability in the second referendum). In the event of a "yes" vote, a cost is randomly selected from a set of higher cost amounts available. For comparability to the SBC treatment, the first cost is a random draw from the vector {\$2, \$3, \$4, \$5} with equal probabilities. To accommodate "yes" responses to the highest cost or "no" responses to the lowest cost, the second cost vector is identical to that used for the SBC treatment.

Research on payment card bid design (Rowe et al., 1996) emphasizes the importance of including a range of bids that sufficiently spans the underlying WTP distribution. The pre-test and the pilot session yield a range of \$0 to \$10, and we set the lowest and highest values of the PC to match these amounts. To allow for comparisons with the SBC and DB formats, the PC includes all integer amounts from \$0 to \$10.

The bid distributions, and sample sizes, are informed by power calculations based on Monte Carlo simulations. The simulations suggest that for a wide range of possible treatment effects (i.e., mean differences between treatments), the PC design yields virtually the same power as an OE elicitation. This is not entirely surprising, as the PC elicits a reasonably tight WTP interval. Mirroring the literature, the simulations confirm that smaller, rather than larger, bid designs are more efficient for a single binary choice elicitation. Nevertheless, these designs lead to considerable power loss when the assumed mean of the distribution is reasonably far away from the true mean. Our bid design is a compromise that, based on the simulations, performs well "on average." The bid distribution we employ has a small but acceptable loss in power relative to the efficient design when the true distribution is known, but performs considerably better across a wide range of possible effect sizes. The simulations suggest that the power of the DB design is nearly as good as the PC and OE elicitations.

We note that two prior controlled experiments on elicitation effects using student samples and public goods find that a single binary choice approach leads to valuations that are roughly twice as large as an open-ended question (Kealy and Turner, 1993; Welsh and Poe, 1998).[11] Our experimental design is powered to detect much smaller treatment effect sizes. In particular, to determine our sample sizes, we set as an objective the detection of a minimum effect size of 70 cents with 80% power. The simulations assume the data are analyzed by an interval regression model (which we in fact use), and that the null hypothesis of equal means is evaluated with a t-test based on a 5% significance level.[12] This requires sample sizes of roughly 100 for each of the OE, PC and DB treatments, and a sample of 130 for the SBC treatment.

[11] Kealy and Turner (1993) estimate that median WTP is 1.4 to 2.7 times larger for the single binary choice format, depending on the model specification. From Welsh and Poe (1998), median WTP from the single binary choice is estimated to be 2.0 times larger.

[12] An effect size of roughly 60 cents can be detected with 80% probability based on a 10% significance level.

2.4 Experimental procedures

A typical session proceeds as follows. Participants are randomly assigned a seat in the lab. The same experiment moderator summarizes lab protocols (e.g., no deception), and reads the experiment instructions aloud while participants follow along on their printed copy. Participants are informed that each is randomly assigned an ID number and that decisions are anonymous. Questions are encouraged by the moderator. The experiment consists of two stages: an earnings stage and a voting stage. All decisions are made on the computer. The experiment is programmed and conducted using the software z-Tree (Fischbacher, 2007).

In the earnings stage, participants earn money by scoring points in two tasks. Both tasks are designed to be "real effort" tasks in the sense that exerting more effort leads to higher earnings, and performance in these tasks is not heavily dependent on cognitive ability or experience. The earnings tasks provide participants with an amount of money more than sufficient to cover the potential cost to them of funding the tree-planting project. Moreover, earnings tasks help to enhance external validity by dampening possible "house money" effects. To determine earnings, scores from the tasks are summed, and all participants are rank-ordered according to their scores. Based on her place in the score distribution, a participant earns either $25 (top 20%), $22.50 (top 21-40%), $20 (top 41-60%), $17.50 (top 61-80%), or $15 (bottom 20%). This procedure induces competition among participants and holds fixed the earnings distribution across sessions.

The first task, developed by Abeler et al. (2011), involves counting the number of zeros in tables that contain randomly generated zeros and ones. Participants have three attempts to enter the correct number of zeros for a given table, after which a new table is generated. For each correctly counted table, a participant earns four points. Five minutes are given for this task. In the second task, developed by Erkal et al. (2011), participants are provided with a table that assigns a number value to each letter of the alphabet, and are asked to encode words into numbers. Participants cannot move to the next word until the currently displayed word is encoded correctly. Similar to the first earnings task, participants have five minutes to encode as many words as possible. For each correctly encoded word, a participant earns one point.[13]

In the next stage of the experiment, that is, in the voting stage, participants are provided information on the tree-planting project. Instructions are identical across treatments, including not only the description of the good, but also the procedure for verifying that money collected will in fact be used to fund the project. The only variation across treatments is the explanation of the elicitation mechanism ("the voting process" in the instructions). After all decisions are collected, DB, PC, and OE participants learn their random individual cost draw. In all treatments, the results screen displays a participant's vote in the binding referendum, the percentage of "yes" and "no" votes in the group, and whether the referendum passes. If the referendum passes, the individual cost of funding the project is subtracted from earnings obtained in the first stage of the experiment. A volunteer is asked to place in the mail a sealed envelope containing a check to GreenTrees along with a letter describing the purpose of the check and that students from the University of Tennessee are funding the project. Upon receipt of payment, participants are emailed a letter from GreenTrees to verify the transaction.

The experiment continues with a short questionnaire that probes participants about their voting decision(s) and collects basic socio-demographic information. The experiment concludes by paying participants their cash earnings in private.

[13] We used results from prior experiments (Abeler et al., 2011; Erkal et al., 2011) to determine conversion rates such that the expected number of points in the two tasks are approximately equal.

2.5 Participants

Four hundred and ten students of the University of Tennessee participated in the experiment from December 2017 to February 2018. All sessions took place in the UT Experimental Economics Laboratory. People were selected from a large pool of students registered as potential participants in economic experiments. The pool resembles the general population of students of the University with respect to gender, age and academic college. Thirty-eight percent of individuals had previously participated in an (unrelated) economics experiment. Participants were not allowed to attend more than one session of the experiment.

Of the 18 sessions conducted, there are six sessions for the SBC treatment and four sessions for the DB, PC, and OE treatments each. The number of participants per session ranged from 16 to 24. A single session lasted about 40 minutes. Average earnings were $19.79, and the referendum passed in seven of the 18 sessions.

3. Data analysis

Table 1 summarizes, by treatment, data obtained from the post-experiment questionnaire. It also includes pre-vote earnings from the first stage of the experiment. The average participant age is 21 years, 43% of the participants are female, 49% are currently employed, and the average reported GPA is 3.27 on a 4-point scale. As anticipated due to random treatment assignment, the distributions of these socio-demographic variables are similar and not statistically different across treatments.

The table further reports summary statistics from a set of three questions included in the questionnaire to gauge how well participants understood experimental procedures. The vast majority of participants (87%) stated that the experiment instructions were overall "well understood." Further, 86% and 75% of participants, respectively, indicated disagreement with the statements "I was confused about the procedure used to determine whether the referendum passed" and "I did not have enough information to make a comfortable decision in the referendum." Using Pearson's chi-squared test of equality of means, we find some statistical differences in the response distributions of these questions across treatments, which point to some distinctions in the comprehension of experimental procedures. Specifically, when compared to the SBC treatment, the PC and OE treatments are characterized by statistically lower comprehension of the instructions (at the 5% significance level for each comparison) and by statistically higher confusion about the voting procedure (at the 1% significance level for the comparison of SBC and PC, and at the 10% significance level for the comparison of SBC and OE). Observing these differences to be significant aligns with expectations, given the higher complexity of the voting procedure in the PC and OE treatments relative to the SBC treatment. Further, relative to DB participants, PC participants appear to have been in need of more information to make a comfortable decision in the referendum (a difference significant at the 10% level). Although statistical differences arise, the magnitudes of the differences highlighted above are small.

We also asked about certainty over a participant's decision in the referendum, as commonly done in stated preference surveys. The mean response of 4.01, on a scale from 1 ("very uncertain") to 5 ("absolutely certain"), suggests the average participant was "certain"; moreover, the modal response was "absolutely certain". The average certainty level is lowest in the OE sample at 3.78, and highest in the DB sample at 4.14. Using Pearson's chi-squared tests, the OE sample distribution of responses to the certainty question is statistically different at the 5% level when compared with the SBC or DB sample. This result is somewhat expected, as the OE elicitation requires participants to provide a point estimate of WTP rather than a "yes" or "no" response to explicit cost levels. No other statistical differences are found in other pairwise comparisons of treatments.

3.1 Nonparametric tests of WTP distributions

Table 2 presents the observed WTP distributions (survival functions) for each treatment. The SBC and PC survival functions are the observed percentages of "yes" votes recorded at each cost amount. The OE survival function is constructed by calculating the percentage of participants indicating a WTP at least as high as a specific cost amount. The survival function for DB data is computed in a similar fashion.[14] The survival functions are monotonically decreasing in all cases.

The survival functions are reasonably similar to one another based on visual inspection. To nonparametrically test for differences in empirical distributions, we use two-sample Kolmogorov-Smirnov (K-S) tests. The test statistic is the absolute value of the largest difference in the observed probabilities across two distributions. A rejection of the null hypothesis can result from differences in the shapes or locations of distributions. The largest observed difference in probabilities across all pairwise comparisons is 0.1286, which occurs at $1 when comparing the DB and PC treatments. The K-S statistics for other comparisons are: 0.0815 (SBC vs. DB); 0.0851 (SBC vs. PC); 0.0798 (SBC vs. OE); 0.0811 (DB vs. OE); and 0.0957 (PC vs. OE). We fail to reject equal distributions in any pairwise comparison.[15] In order to estimate mean WTP values, and to further condition them on participants' socio-demographic characteristics and on the money earned from the pre-vote tasks, we now turn to econometric modelling.

[14] For the DB survival function, when calculating the percentage of "yes" votes at $x, we omit observations for which $x falls within the elicited WTP interval. For example, if a participant votes "yes" to $2 and "no" to $5, this is interpreted as a "yes" vote to $1 and $2, missing values for $3 and $4, and a "no" vote to $5 and $6.

[15] Critical values for the K-S test depend on sample sizes. For these comparisons, the 5% critical values are in the 0.18 to 0.20 range, and 10% critical values span from 0.16 to 0.18.

3.2 Econometric analysis

The data generated from the four value elicitation mechanisms provide continuous, left-censored, right-censored or interval-censored signals of participants' WTP. The maximum likelihood estimator we specify, which nests the estimators of Cameron and James (1987) and Cameron and Huppert (1989), accommodates the information obtained across the elicitation formats in a unified way, enabling tests of treatment effects that are not driven by differences in statistical assumptions. The estimation allows for the possibility that some participants have negative WTP, as suggested by the finding that 17% voted "no" at a cost of $0 in the PC treatment. Further, 21% of the OE sample indicated $0 WTP, which may also signal negative WTP (noting that negative valuations were not permissible).

Let WTP_i denote participant i's willingness to pay for the project. WTP_i is not directly observed, aside from most OE participants, but instead can be treated as a censored dependent variable. For the PC elicitation, we obtain the signal $c_{i,l} \leq WTP_i < c_{i,u}$, where $c_{i,l}$ is the highest cost for which participant i votes "yes" and $c_{i,u}$ is the next higher amount. For the case where the participant votes "no" to the lowest amount, $c_{i,l} = -\infty$ and $c_{i,u}$ is equal to the lowest amount; similarly, $c_{i,l}$ is equal to the highest amount and $c_{i,u} = +\infty$ if she votes "yes" to the highest amount. For the SBC elicitation, we obtain the signal $WTP_i < c_i$ if the participant votes "no" to the stated cost c_i, and the signal $WTP_i \geq c_i$ for a "yes" vote. As such, SBC data represents a special case of PC data, where c_i defines only an upper or lower bound (e.g., $c_{i,l} = c_i$ and $c_{i,u} = +\infty$ for a "yes" vote). For DB responses, a well-defined interval emerges in cases where a "yes" vote is observed for just one of the two votes. Otherwise, for two "no" votes, the data is left-censored (that is, $c_{i,l} = -\infty$) and the lower amount offered forms the upper bound $c_{i,u}$. For two "yes" votes, the higher of the two amounts offered forms the lower bound $c_{i,l}$ and the data is right-censored (that is, $c_{i,u} = +\infty$). Finally, for OE responses, WTP_i is directly observed, with the exception of zero valuations, which we interpret to be left-censored to allow for negative WTP, which is consistent with our treatment of the other data types. These left-censored observations are accommodated by defining a WTP interval with $c_{i,l} = -\infty$ and $c_{i,u} = 0$.

Assume WTP_i is a linear function of a row vector of covariates, $\mathbf{x}_i$, such that $WTP_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$, where $\boldsymbol{\beta}$ is a column vector of unknown parameters and ε_i is a normally distributed mean-zero error term with standard deviation σ_i. With the linear conditional mean function, assuming the error term has a normal distribution is analogous to assuming a normal distribution for WTP_i. Moreover, interpretation of estimated parameters is the same as for a standard linear regression model that treats WTP_i as a directly observed (i.e., uncensored) dependent variable. Let $D_i = 1$ if the response is censored (that is, SBC, DB, PC, and zero OE responses), and $D_i = 0$ if the response is uncensored (that is, OE responses higher than zero). Then, the log-likelihood function for the WTP regression model is

$$\ln \mathcal{L} = \sum_{i=1}^{N} \left\{ D_i \cdot \ln\left(\Phi\left(\frac{c_{i,u} - \mathbf{x}_i\boldsymbol{\beta}}{\sigma_i}\right) - \Phi\left(\frac{c_{i,l} - \mathbf{x}_i\boldsymbol{\beta}}{\sigma_i}\right) \right) + (1 - D_i) \cdot \ln\left(\frac{1}{\sigma_i} \phi\left(\frac{WTP_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma_i}\right) \right) \right\},$$

where Φ and ϕ denote the CDF and PDF of the standard normal distribution, respectively. The first term corresponds with the log-likelihood for (interval) censored data, whereas the second term corresponds to that of a normal regression model for uncensored data.[16] The OE data are a mix of censored and uncensored data, and applying the above estimator is equivalent to a using Tobit with left-censoring at zero.

[16] Given how we code the upper and lower bounds, the contribution to the log-likelihood for a SBC participant is mathematically equivalent to that of a probit model.

Studies such as Haab et al. (1999) highlight the importance of allowing for different error variances when pooling preference data from different experimental treatments. To allow for possibly different error variances across elicitation formats, we define $\sigma_i = \sigma_0 + \sigma_1 DB_i + \sigma_2 PC_i + \sigma_3 OE$, where OE_i, DB_i, and PC_i are treatment-specific indicator variables that equal 1 for OE, DB, and PC observations, respectively. With this specification, σ_0 is the standard deviation of the error term for the SBC data.

Table 3 reports the results of the WTP regressions. Model I allows the mean WTP to vary across treatments, but constrains the error variances to be equal. Model II extends the specification to allow for unequal variances. Model III additionally includes variables that control for participants' socio-demographic characteristics and the amount of earned income in the pre-vote experiment tasks. These variables may explain some of the variation in WTP across participants, while also adjusting estimates for unintended differences due to sampling.

Considering all three models, there are no statistical differences in mean WTP across treatments. Mean WTP falls in the \$3.66 (Model II, OE treatment) to the \$4.00 (Model III, PC treatment) range. Thus, point estimates vary by less than 10% across all models and treatments. Models II and III further demonstrate that the variances of the treatment-specific WTP distributions are statistically equal. This is overall very strong evidence of convergent validity, and corroborates the findings from the K-S tests.[17] Model III suggests that being older and female is positively correlated with WTP for the tree-planting project. There is a negative correlation between participation in a previous, unrelated economics experiment, and this correlation is marginally significant. This result is consistent with evidence from prior research, which suggests

[17] For Models II and III, we also fail to reject the null hypothesis that both the mean and standard deviation of the WTP distribution are statistically equal across treatments, for all pairwise comparisons.

that experienced subjects are more strongly motivated by the maximization of their own earnings (see Schmidt et al., 2018).

Evidence from field surveys suggests that the DB format may invoke behavioral responses or updating that leads to differences in the mean or variance of WTP across the value elicitation questions. To explore this issue, following Cameron and Quiggin (1994), we estimate a bivariate normal model for interval-censored data, allowing for different means and variances across the two DB valuation questions. This results in a mean WTP of $4.03 (std. err. = 0.32) for the first question and $3.53 (0.75) for the second. The estimated standard deviations are 2.20 (0.60) and 5.88 (2.47), respectively, and the correlation coefficient is 0.90 (0.10). Both the means and standard deviations are statistically equal across the two questions (*p*-value = 0.33).

4. Lab-to-field generalizability

4.1 Generalizability of results

In retrospect, our unexpected findings may have been more powerful had we also tested for elicitation effects using stated preference methods.[18] Were we able to replicate the stylized fact in the literature, this would help dispel possible concerns that our findings may be an artifact of our particular subject pool and experimental methods. Absent this evidence, in this section we use insights from the stated preference and broader experimental economics literatures to speak to the generalizability of our findings, which of course remains an open empirical question.

A prominent concern with laboratory experiments such as ours is the use of college student participants. The accumulated evidence from broader experimental economics research suggests, however, that important differences in treatment effects between students and participants drawn

[18] We thank a careful reviewer of this manuscript for emphasizing this point.

from other populations are rare (Fréchette, 2016). As articulated by Cason and Wu (2019), experiments with targeted populations are important when the objective is measurement, but qualitative findings are unlikely to be affected by the subject pool. Stated another way, we do *not* expect mean WTP estimates from our experiment to be robust to other populations (even to students at a different university), but we do expect our finding of no elicitation effects to be robust. Drawing specifically from the stated preference literature, we note that some of the most prominent findings, such as the WTP and willingness to accept (WTA) disparity (Tunçel and Hammitt, 2014), and differences between real and hypothetical choices (Penn and Hu, 2018), arise in both student and non-student experiments.

For more convincing evidence, we searched the literature for stated preference studies of elicitation effects that involved public goods and were conducted in a controlled setting with college students. Although we only uncovered two studies that meet these criteria, both replicate the common finding that a single binary choice format yields substantially higher WTP estimates (Kealy and Turner, 1993; Welsh and Poe, 1998). For example, estimates from Welsh and Poe (1998) suggest that median WTP based on a single binary choice question is 2.0 times larger than that elicited from an open-ended question, and 2.9 times larger than from a payment card. Similar to our study, elicitation formats in their experiment are framed as referenda, the payment vehicle is well-defined, and the good participants are asked to value is clearly described.

A related concern is whether our results are robust to changing the procedure used to provide participants with income. On one extreme, we could have simply gifted participants with money. On the other, we could have provided no income, leaving participants to pay out of pocket for tree plantings. Evidence from dictator games and charitable giving experiments suggests that when income is earned, either from a laboratory task or brought from outside the lab, people behave more selfishly relative to when income is gifted (see Reinstein and Reiner, 2012). From this, we would naturally expect that providing unearned income would increase WTP in our study. However, there is limited evidence on whether there are important links between the method for providing income and *treatment* effects. Cherry et al. (2005) find that, in a linear public goods game, contribution levels decrease with income heterogeneity regardless of the origin of income (unearned or earned). As a more closely related example, accumulated evidence suggests that providing participants with an endowment or not does not alter findings of hypothetical bias (Penn and Hu, 2018).

4.2 Applicability of experimental methods to the field

A single binary choice question is commonly viewed as providing the strongest incentive for respondents to truthfully reveal preferences (Johnston et al., 2017). Nevertheless, alternative formats carry advantages such as statistical efficiency and convenience. For example, an open-ended format elicits a point estimate of WTP, and prior knowledge of the population WTP distribution is not necessary. In contrast, the single binary choice approach only elicits an upper or lower bound on WTP for each respondent, and the efficiency of the bid design depends critically on knowledge of the population WTP distribution (Cooper, 1993). Improving the incentive properties of alternative formats is thus an important objective, and below we discuss how our experimental methods may translate to field settings.

Our experimental value elicitation methods rely on provision rules and direct financial incentives, neither of which can be precisely applied in most field survey studies. Several characteristics of the experimental design can nevertheless be mirrored in surveys to a large degree, with the potential benefit of mitigating elicitation effects. Importantly, these characteristics are congruent with the recommendations of Johnston et al. (2017), who state that "incentive-compatible response formats are preferred" (p. 345), "referendum formats should be considered when plausible" (p. 351), and the "payment vehicle should be selected to be…binding for all respondents" (p. 352).

All our elicitations are framed as referenda and incorporate a coercive payment vehicle. Aside from single binary choice elicitations, it is uncommon for field surveys to include one let alone both of these features. Our DB format is characterized by "advanced disclosure" in the sense that, prior to voting, participants are told they will participate in two referenda. Similar forms of disclosure have been shown to reduce behavioral anomalies in stated preference surveys (Bateman et al., 2004; Day et al., 2012), although this approach is infrequently applied. Last, our provision rules induce uncertainty in order to break possible strategic links that would otherwise lead to a loss of incentive compatibility.

Inducing uncertainty in a field survey is likely the most challenging. Nevertheless, Vossler and Holladay (2018) demonstrate that cost uncertainty can be translated into field elicitations. In that study, the authors implement an open-ended elicitation that informs households that the cost of the project is unknown (indeed, costs of a large-scale project are rarely *ex ante* known with certainty), and that results from the elicitation can be used to understand the percentage in favor of the project in the event the actual cost becomes known. This language captures our OE treatment, at least in principle. In survey settings where multiple goods are being valued, such as in a DCE, inducing uncertainty over possible policies is theoretically desirable (Vossler et al., 2012). This can be operationalized by modifying the survey scripts of Vossler and Holladay (2018) to focus on policy uncertainty rather than cost uncertainty.

One concern over introducing uncertainty in survey scripts is that respondents may not find this to be credible, possibly giving rise to considerable bias in elicited values. While more evidence on this is needed, the results from Vossler and Holladay (2018) suggest optimism. The authors report that no issues arose in focus groups regarding scenario credibility. Further, when comparing two open-ended elicitations – a theory-driven one with cost uncertainty and a standard one without cost uncertainty – they find that respondents to the former are *more* likely to view the survey as both payment and policy consequential.

On a final note, while the survey of Vossler and Holladay (2018) serves as the primary example of how characteristics of our experimental methods may be captured in field settings, we believe there is one important source of procedural variance in their study: their payment card and open-ended surveys suggest that costs would be the same for all households. In our experiment, we left ambiguous whether different participants might pay different amounts upon implementation (in reality, payments do differ), which is consistent with single binary choice surveys.

5. Conclusion

In this study, we test the importance of behavioral factors, such as anchoring and "yea-saying", in explaining the well-established stylized fact in the stated preference literature that different question response formats lead to different estimates of economic welfare. Our approach relies on holding fixed economic incentives, which have also been hypothesized to explain the so-called elicitation effects, across four elicitation formats: single binary choice, double-bounded binary choice, payment card, and open-ended. We elicit homegrown values for an environmental public good. Reflecting common field applications of contingent valuation surveys, values for the good are largely based on passive use, and participants are unlikely to have prior experience in placing values on the good or related goods.

Given the importance attributed to behavioral factors not only in the stated preference literature, but also in the broader economics literature, our *a priori* expectation was that we would find elicitation effects similar to what has been found in the field. To our surprise, our key result instead is that the four elicitation formats examined lead to statistically identical willingness-to-pay (WTP) distributions. This result suggests that – at least for a subset of elicitation formats – differences in perceived economic incentives in the field setting may be the primary driver of elicitation effects. While the methods we employ to control these incentives cannot seamlessly be translated to the field, we make several recommendations on survey design in the preceding section.

While we did not find any evidence that behavioral factors drive a wedge between formats in our experiment, behavioral factors may nevertheless be important. It is plausible that behavioral drivers become prominent when economic incentives are either weak or poorly understood; i.e., there may exist important interaction effects. Such behavior would be predicted by the model of Smith and Walker (1993), which postulates that people weigh decision costs (i.e., cognitive burden) and monetary rewards, and that decisions are more likely to adhere to rational choice models as rewards increase.

The literature suggests ways in which our experimental design can be extended to test whether behavioral drivers emerge as economic incentives are weakened, or otherwise to better reflect field conditions. For instance, Carson et al. (2014) vary the probability that a vote is binding. Vossler et al. (2012) frame the elicitation as a vote but keep the decision rule undisclosed. In a future test of the double-bounded format, for example, one possibility is to have participants vote “yes” or “no” to two possible costs, but provide no information on how this information will map into a binding outcome. These investigations, in turn, may identify a subset of elicitation formats and conditions under which elicitation effects do not to emerge. Accumulated evidence suggests that contingent valuation surveys match well the outcomes of binding, public referenda in settings where single binary choice survey elicitations are believed to adhere to incentive compatibility assumptions (e.g., Johnston, 2006). It follows logically that if researchers uncover design modifications that mitigate elicitation effects, this in turn should enhance the criterion validity of other formats. This is to say that results from the lab are likely to be useful in informing field studies.

On a final note, our examination is limited to four elicitation formats, whereas many others are used in practice (see, for example, Carson and Louviere, 2011). The variation in the formats we investigate coincides well with key characteristics of extant formats: the precision at which values are elicited and the number of value questions used. Our research involves both open-ended and close-ended formats, considers both single and repeated choice valuation formats, and includes repeated formats that span sequential (i.e., double-bounded) and simultaneous (i.e., payment card) decision settings. Nevertheless, there is merit in using controlled experiments with field context to study other important formats, especially those that are shown to reveal demand in an induced-value setting.

References

Abeler, Johannes, Armin Falk, Lorenz Goette, and David Human. 2011. Reference points and effort provision. *American Economic Review* 101(2): 470-492.

Alberini, Anna. 1995. Testing willingness-to-pay models of discrete choice contingent valuation survey data. *Land Economics* 71(1): 83-95.

Alevy, Jonathan E., Craig E. Landry, and John A. List. 2015. Field experiments on the anchoring of economic valuations. *Economic Inquiry* 53(3): 1522-1538.

Ami, Dominique, Frédéric Aprahamian, Olivier Chanel, and Stéphane Luchini. 2011. A test of cheap talk in different hypothetical contexts: The case of air pollution. *Environmental and Resource Economics* 50(1): 111-130.

Attiyeh, Greg, Robert Franciosi, and R. Mark Isaac. 2000. Experiments with the pivot process for providing public goods. *Public Choice* 102(1-2): 93-112.

Azrieli, Yaron, Christopher P. Chambers, and Paul J. Healy. 2018. Incentives in experiments: A theoretical analysis. *Journal of Political Economy* 126(4): 1472-1503.

Bateman, Ian J., Matthew Cole, Philip Cooper, Stavros Georgiou, David Hadley, and Gregory L. Poe. 2004. On visible choice sets and scope sensitivity. *Journal of Environmental Economics and Management* 47(1): 71-93.

Bateman, Ian J., Philip Cooper, Stavros Georgiou, Ståle Navrud, Gregory L. Poe, Richard C. Ready, Pere Reira, Mandy Ryan, and Christian A. Vossler. 2005. Economic valuation of policies for reducing acidity in remote mountain lakes. *Aquatic Sciences* 67(3): 274-291.

Bateman, Ian J., Ian H. Langford, and Jon Rasbash. 2001. Willingness-to-pay question format effects in contingent valuation studies. In Ian J. Bateman and Kenneth G. Willis (eds.), *Valuing environmental preferences: Theory and practice of the contingent valuation method in the US, EU, and developing countries*. Oxford: Oxford University Press.

Becker, Gordon M., Morris H. DeGroot, and Jacob Marschak. 1964. Measuring utility by a single-response sequential method. *Behavioral Science* 9(3): 226-232.

Bishop, Richard C., Kevin J. Boyle, Richard T. Carson, David Chapman, W. Michael Hanemann, Barbara Kanninen, Raymond J. Kopp, Jon Krosnick, John List, Norman Meade, Robert Paterson, Stanley Presser, V. Kerry Smith, Roger Tourangeau, Michael Welsh, Jeffrey M. Wooldridge, Matthew De Bell, Colleen Donovan, Matthew Konopka, and Nora Scherer. 2017. Contingent valuation: Flawed logic? – Response. *Science* 357(6349): 363-364.

Boyle, Kevin J. 1989. Commodity specification and the framing of contingent-valuation questions. *Land Economics* 65(1): 57-63.

Cameron, Trudy A., and Daniel D. Huppert. 1989. OLS versus ML estimation of non-market resource values with payment card interval data. *Journal of Environmental Economics and Management* 17(3): 230-246.

Cameron, Trudy A., and Michael D. James. 1987. Efficient estimation methods for use with "closed-ended" contingent valuation survey data. *The Review of Economics and Statistics* 69(2): 269-276.

Cameron, Trudy. A., Gregory L. Poe, Robert G. Ethier, and William D. Schulze. 2002. Alternative non-market value-elicitation methods: Are the underlying preferences the same? *Journal of Environmental Economics and Management* 44(3): 391-425.

Cameron, Trudy A., and John Quiggin. 1994. Estimation using contingent valuation data from a "dichotomous choice with follow-up" questionnaire. *Journal of Environmental Economics and Management* 27(3): 218-234.

Carson, Katherine Silz, Susan M. Chilton, and W. George Hutchinson. 2009. Necessary conditions for demand revelation in double referenda. *Journal of Environmental Economics and Management* 57(2): 219-255.

Carson, Richard T., and Theodore Groves. 2007. Incentive and informational properties of preference questions. *Environmental and Resource Economics* 37(1): 181-210.

Carson, Richard T., Theodore Groves, and John A. List. 2014. Consequentiality: A theoretical and experimental exploration of a single binary choice. *Journal of the Association of Environmental and Resource Economists* 1(1-2): 171-207.

Carson, Richard T., and Jordan J. Louviere. 2011. A common nomenclature for stated preference elicitation approaches. *Environmental and Resource Economics* 49(4): 539-559.

Cason, Timothy N., and Steven Y. Wu. 2019. Subject pools and deception in agricultural and resource economics experiments. *Environmental and Resource Economics* 73: 743-758.

Champ, Patricia A., and Richard C. Bishop. 2006. Is willingness to pay for a public good sensitive to the elicitation format? *Land Economics* 82(2): 162-173.

Champ, Patricia A., Nicholas E. Flores, Thomas C. Brown, and James Chivers. 2002. Contingent valuation and incentives. *Land Economics* 78(4): 591-604.

Cherry, Todd L., Stephan Kroll, and Jason F. Shogren. 2005. The impact of endowment heterogeneity and origin on public good contributions. *Journal of Economic Behavior and Organization* 57(3), 357-365.

Collins, Jill P., and Christian A. Vossler. 2009. Incentive compatibility tests of choice experiment value elicitation questions. *Journal of Environmental Economics and Management* 58(2): 226-235.

Cooper, David J., and Hanming Fang. 2008. Understanding overbidding in second price auctions: An experimental study. *Economic Journal* 118: 1572-1595.

Cooper, Joseph C. 1993. Optimal bid selection for dichotomous choice contingent valuation surveys. *Journal of Environmental Economics and Management* 24: 25-40.

Cox, James C., Vjollca Sadiraj, and Ulrich Schmidt. 2015. Paradoxes and mechanisms for choice under risk. *Experimental Economics* 18(2): 215-250.

Day, Brett, Ian J. Bateman, Richard T. Carson, Diane Dupont, Jordan J. Louviere, Sanae Morimoto, Riccardo Scarpa, and Paul Wang. 2012. Ordering effects and choice set awareness in repeat-response stated preference studies. *Journal of Environmental Economics and Management* 63(1): 73-91.

Erkal, Nisvan, Lata Gangadharan, and Nikos Nikiforakis. 2011. Relative earnings and giving in a real-effort experiment. *American Economic Review* 101(7): 3330-3348.

Farquharson, Robin. 1969. *Theory of Voting*. New Haven, CT: Yale University Press.

Fischbacher, Urs. 2007. Zurich toolbox for ready-made economic experiments. *Experimental Economics* 10(2): 171-178.

Fréchette, Guillaume R. 2016. Experimental economics across subject populations. In John H. Kagel and Alvin E. Roth (eds.), *The Handbook of Experimental Economics, vol. 2*. Princeton, NJ: Princeton University Press.

Frew, Emma J., David K. Whynes, and Jane L. Wolstenholme. 2003. Eliciting willingness to pay: Comparing closed-ended with open-ended and payment scale formats. *Medical Decision Making* 23: 150-159.

Fundenberg, Drew, David K. Levine, and Zacharias Maniadis. 2012. On the robustness of anchoring effects in WTP and WTA experiments. *American Economic Journal: Microeconomics* 4(2): 131-145.

Green, Donald, Karen E. Jacowitz, Daniel Kahneman, and Daniel McFadden. 1998. Referendum contingent valuation, anchoring, and willingness to pay for public goods. *Resource and Energy Economics* 20(2): 85-116.

Haab, Timothy C., Ju-Chin Huang, and John C. Whitehead. 1999. Are hypothetical referenda incentive compatible? A comment. *Journal of Political Economy* 107(1): 186-196.

Hanemann, W. Michael. 1995. Contingent Valuation and Economics. In K.G. Willis and J.T. Corkindale (eds.), *Environmental Valuation New Perspectives*. Oxon: CAB International.

Huang, Ju Chin, and V. Kerry Smith. 1998. Monte Carlo benchmarks for discrete response valuation methods. *Land Economics* 74(2): 186-202.

Johnston, Robert J. 2006. Is hypothetical bias universal? Validating contingent valuation responses using a binding public referendum. *Journal of Environmental Economics and Management* 52(1): 469-481.

Johnston, Robert J., Kevin J. Boyle, Wiktor L. Adamowicz, Jeff Bennett, Roy Brouwer, Trudy A. Cameron, W. Michael Hanemann, Nick Hanley, Mandy Ryan, Riccardo Scarpa, Roger Tourangeau, and Christian A. Vossler. 2017. Contemporary guidance for stated preference studies. *Journal of the Association of Environmental and Resource Economists* 4(2): 319-405.

Kanninen, Barbara J. 1995. Bias in discrete response contingent valuation. *Journal of Environmental Economics and Management* 28(1): 114-125

Kealy, Mary Jo, and Robert W. Turner. 1993. A Test of the Equality of Closed-Ended and Open-Ended Contingent Valuations. *American Journal of Agricultural Economics* 75(2): 321-331.

Kling, Catherine L., Daniel J. Phaneuf, and Jinhua Zhao. 2012. From Exxon to BP: Has some number become better than no number? *Journal of Economic Perspectives* 26(4): 3-26.

McFadden, Daniel, and Gregory K. Leonard. 1993. Issues in the Contingent Valuation of Environmental Goods. In Jerry A. Hausman (ed.), *Contingent Valuation: A Critical Assessment.* Amsterdam, The Netherlands: North Holland Publishing Co. (Elsevier).

Messer, Kent D., Gregory L. Poe, Daniel Rondeau, William D. Schulze, and Christian A. Vossler. 2010. Social preferences and voting: An exploration using a novel preference revealing mechanism. *Journal of Public Economics* 94(3-4): 308-317.

O'Conor, Richard M., Magnus Johannesson, and Per-Olov Johansson. 1999. Stated preferences, real behaviour and anchoring: Some empirical evidence. *Environmental and Resource Economics* 13(2): 235-248.

Penn, Jerrod M., and Wuyang Hu. 2018. Understanding hypothetical bias: An enhanced meta-analysis. *American Journal of Agricultural Economics* 100(4): 1186-1206.

Plott, Charles R., and Kathryn Zeiler. 2005. The willingness to pay – willingness to accept gap, the "endowment effect," subject misconceptions, and experimental procedures for eliciting valuations. *American Economic Review* 95(3): 530-545.

Poe, Gregory L. 2016. Behavioral anomalies in contingent values and actual choices. *Agricultural and Resource Economics Review* 45(2): 246-269.

Ready, Richard C., Ståle Navrud, and W. Richard Dubourg. 2001. How do respondents with uncertain willingness to pay answer contingent valuation questions? *Land Economics* 77(3): 315-326.

Reinstein, David, and Gerhard Riener. 2012. Decomposing desert and tangibility effects in a charitable giving experiment. *Experimental Economics* 15(1): 229-240.

Rowe, Robert D., William D. Schulze, and William S. Breffle. 1996. A test for payment card biases. *Journal of Environmental Economics and Management* 31(2): 178-185.

Rutström, E. Elisabet. 1998. Home-grown values and the design of incentive compatible auctions. *International Journal of Game Theory* 27: 427-441.

Scheufele, Gabriela, and Jeff Bennett. 2013. Effects of alternative elicitation formats in discrete choice experiments. *Australian Journal of Agricultural and Resource Economics* 57(2): 214-233.

Schmidt, Robert J., Christiane Schwieren, and Alec N. Sproten. 2018. Social norm perception in economic laboratory experiments: Inexperienced versus experienced participants. University of Heidelberg, Department of Economics, Discussion Paper Series No. 656.

Smith, Vernon L., and James M. Walker. 1993. Monetary rewards and decision cost in experimental economics. *Economic Inquiry* 31(2): 245-261.

Taylor, Laura O., Michael McKee, Susan K. Laury, and Ronald G. Cummings. 2001. Induced-value tests of the referendum voting mechanism. *Economics Letters* 71(1): 61-65.

Tunçel, Tuba, and James K. Hammitt. 2014. A new meta-analysis on the WTP/WTA disparity. *Journal of Environmental Economics and Management* 68(1): 175-187.

Vossler, Christian A., Maurice Doyon, and Daniel Rondeau. 2012. Truth in consequences: Theory and field evidence on discrete choice experiments. *American Economic Journal: Microeconomics* 4(4): 145-171.

Vossler, Christian A., and J. Scott Holladay. 2018. Alternative value elicitation formats in contingent valuation: Mechanism design and convergent validity. *Journal of Public Economics* 165: 133-145.

Vossler, Christian A., and Michael McKee. 2006. Induced-value tests of contingent valuation elicitation mechanisms. *Environmental and Resource Economics* 35(2): 137-168.

Welsh, Michael P., and Gregory L. Poe. 1998. Elicitation effects in contingent valuation: Comparisons to a multiple bounded discrete choice approach. *Journal of Environmental Economics and Management* 36(2): 170-185.

Whitehead, John C. 2002. Incentive incompatibility and starting-point bias in iterative valuation questions. *Land Economics* 78(2): 285-297.

Table 1. Summary Statistics by Treatment

	Single Binary Choice (SBC)	Double-Bounded Binary Choice (DB)	Payment Card (PC)	Open-Ended (OE)
Age	20.65	20.80	20.53	20.79
	(3.31)	(2.79)	(2.28)	(1.51)
Female	0.45	0.41	0.37	0.48
	(0.50)	(0.50)	(0.49)	(0.50)
Earned Income	19.77	19.84	19.79	19.79
	(3.54)	(3.49)	(3.49)	(3.49)
Employed	0.46	0.58	0.47	0.48
	(0.50)	(0.50)	(0.50)	(0.50)
GPA	3.19	3.34	3.22	3.36
	(0.57)	(0.50)	(0.50)	(0.43)
Comprehension	4.90	4.88	4.79	4.79
	(0.30)	(0.36)	(0.55)	(0.48)
Confusion	1.35	1.51	1.70	1.54
	(0.73)	(0.88)	(1.12)	(0.99)
Need Information	1.89	1.77	2.10	1.96
	(1.12)	(0.89)	(1.31)	(1.15)
Certainty	4.08	4.14	4.02	3.78
	(0.99)	(0.91)	(1.14)	(1.13)
N	130	92	94	94

Notes: Standard deviations are reported in parentheses. *Earned Income* corresponds with experiment earnings obtained prior to the voting task. GPA is measured on a 4-point scale. *Comprehension* is a 1 ("poorly understood") to 5 ("well understood") indication of instruction clarity. *Confusion* is a 1 ("completely disagree") to 5 ("completely agree") indication of whether participant was confused about the voting process. *Need Information* is a 1 ("completely disagree") to 5 ("completely agree") indication of whether participant had enough information to make a comfortable decision in the referendum. *Certainty* is a 1 ("very uncertain") to 5 ("absolutely certain") indication of a participant's certainty about her voting decision(s).

Table 2. Empirical Survival Functions – Percentage of "Yes" Votes by Cost

Cost	Single Binary Choice (SBC)	Double-Bounded Binary Choice (DB)	Payment Card (PC)	Open-Ended (OE)
$0			82.98	
$1	79.17	87.32	74.47	84.04
$2	72.73	75.00	67.02	71.28
$3	61.90	56.58	56.38	59.57
$4	50.00	50.67	41.49	42.55
$5	33.33	31.94	36.17	35.11
$6	25.00	20.55	20.21	17.02
$7			17.02	13.83
$8			12.77	9.58
$9			12.77	8.51
$10			12.77	8.51

Table 3. Willingness-to-Pay Regressions

	I	II	III
Double-Bounded Binary Choice (DB)	-0.10	0.00	-0.03
	(0.68)	(0.62)	(0.64)
Payment Card (PC)	-0.13	-0.02	0.20
	(0.65)	(0.56)	(0.56)
Open-Ended (OE)	-0.25	-0.18	-0.12
	(0.65)	(0.60)	(0.64)
Age			0.30^{***}
			(0.10)
Female			1.07^{**}
			(0.45)
Earned Income			-0.06
			(0.06)
Employed			0.18
			(0.44)
GPA			0.57
			(0.43)
Participant in Prior Experiment			-0.90^{*}
			(0.49)
Intercept	3.94^{***}	3.84^{***}	3.80^{***}
	(0.48)	(0.38)	(0.41)
Standard deviation function (σ)			
Double-Bounded Binary Choice (DB)		0.89	0.60
		(0.99)	(1.04)
Payment Card (PC)		0.65	0.18
		(0.81)	(0.89)
Open-Ended (OE)		1.24	1.06
		(0.81)	(0.89)
Constant	4.15^{***}	3.22^{***}	3.42^{***}
	(0.23)	(0.73)	(0.82)
Log-*L*	-669.13	-667.92	-657.81
N	410	410	410

Notes: All socio-demographic variables are demeaned so that the Intercept can be interpreted as the estimated mean WTP for the SBC treatment in all models. *** denotes significance at the 1% level, ** significance at the 5% level, and * significance at the 10% level.