Journal of Choice Modelling 29 (2018) 17–32

Contents lists available at ScienceDirect

Journal of Choice Modelling

journal homepage: www.elsevier.com/locate/jocm

Testing the effect of changes in elicitation format, payment vehicle and bid range on the hypothetical bias for moral goods

Lea S. Svenningsen[a,*], Jette Bredahl Jacobsen[a,b]

[a] *Department of Food and Resource Economics, University of Copenhagen, Denmark*
[b] *Centre for Macroecology, Evolution and Climate, University of Copenhagen, Denmark*

ARTICLE INFO

Keywords:
Choice experiment
Hypothetical bias
bid range
Elicitation format
Payment vehicle
Learning effects
Public good

ABSTRACT

This paper explores how changes in survey design influence the conclusions reached from discrete choice models, a topic which is of particular interest in the context of stated and revealed preference comparisons investigating potential hypothetical bias. We systematically test the WTP of a good with no related market value, using two standard, hypothetical stated preference data collections and an incentivized stated preference data collection, using a real donation mechanism. The investigations into the nature of hypothetical bias typically involve changes in more than just the elicitation format. Therefore, we explicitly test the importance of changes in bid range, payment vehicle, and elicitation format upon the estimated hypothetical bias, while keeping the survey context constant. Our results show that depending on the characteristics of the good in question, the choice of payment vehicle, bid range and elicitation format may affect the results. In many cases, the importance of payment vehicle itself is negligible – especially when the good in question is distant to people. The implication of our study is that caution should be applied when conducting stated and revealed preference comparisons in the context of public good with strong moral components, as even very small design decisions may influence the observed WTP disparities.

1. Introduction

The valuation of non-market goods through the use of stated preference techniques is an established practice, allowing policymakers to infer the value placed on goods such as environmental services when no value can be inferred from market goods and related behaviour. Since the emergence and broad appliance of stated preference methods, many rich scientific discussions have taken place regarding the merits and drawbacks of the method (Kahneman and Knetsch, 1992; Carson and Mitchell, 1995; Carson et al., 1996; Johnston et al., 2017). The most prominent factor identified in this debate is the discussion regarding *hypothetical bias*, a term used to capture the idea that when faced with no real consequence of their choices, people will tend to overstate their willingness to pay for a certain good (List and Gallet, 2001; Murphy et al., 2005; Hensher, 2010). While for some types of goods it is possible to triangulate the results, for many other goods it is not. These other goods are typically described by a strong non-use component, also often involving morally challenging decisions. This can for example be conserving nature, or ensuring future generations' living conditions – issues that are indeed discussed and taken into account in policy making. Consequently, if we want to inform policymakers on issues pertinent to policy design, stated preference methods must be applied, despite the existence of this possible bias.

[*] Corresponding author.
E-mail address: leasvenningsen@ifro.ku.dk (L.S. Svenningsen).

https://doi.org/10.1016/j.jocm.2018.08.001
Received 4 December 2017; Received in revised form 9 August 2018; Accepted 12 August 2018
Available online 13 August 2018
1755-5345/

In the literature, the origin of hypothetical bias has been ascribed to several factors. Examples of such factors are the elicitation format and the elicited welfare measure, e.g. willingness-to-pay (WTP) or willingness-to-accept (WTA) (List and Gallet, 2001), the warm glow of giving as termed by Andreoni (1990) which has been subsequently tested in relation to hypothetical bias in stated preference studies (Nunes and Schokkaert, 2003; Nunes et al., 2009; Johansson-Stenman and Svedsäter, 2012), social desirability bias whereby individuals' valuation of a certain good is influenced by how socially attractive it is to show support (Crowne and Marlowe, 1960; Epley and Dunning, 2000; List et al., 2004; Lusk and Norwood, 2009), and lastly also the type of good being valued, e.g. public or private good (Alpizar et al., 2008; Johansson-Stenman and Svedsäter, 2012). In this study, we explore the role of elicitation format, payment vehicle and bid range, which are all factors that may influence the motivational reasons for hypothetical bias.

We use the term elicitation format to capture whether preferences are elicited through a hypothetical or a revealed preference mechanism. The typical approach used in investigating hypothetical bias has been to compare preference measures from stated and revealed preference (SP-RP) counterparts (List and Gallet, 2001). In this study, we are interested in investigating distributional preferences for climate policies, where no revealed preference measure exists. Our survey design scheme facilitates a partial SP-RP preference comparison, as we elicit preferences and welfare measures for climate policy in both two strictly hypothetical settings and in an incentivized stated preference setting, where individuals are making real donations to the same policies as presented to respondents in the hypothetical settings. This change in elicitation format is the closest approximation to a true stated and revealed preference comparison that is possible in the given context, as the most realistic payment vehicle for the climate policies investigated is a tax, which is a payment vehicle that is difficult to reproduce in a revealed preference context.[1] However, in the survey design scheme examined in this paper, not only the elicitation format changes but also the type of payment vehicle and bid range of the cost attribute change, both of which could influence hypothetical bias. Therefore we need to look at the findings in the literature regarding the role of these two additional factors.

A branch of the valuation literature has investigated how a change in the type of payment vehicle in itself might influence inference from choice data, with some papers suggesting that a coercive instrument such as a tax results in higher estimates of willingness-to-pay compared to a voluntary instrument like a donation (Nunes and Schokkaert, 2003; Wiser, 2007). On the other hand, using an instrument like a donation mechanism has been shown to lead to warm glow of giving. While this is indeed a part of the people's utility function, it may not be solely linked to the good in question. As there is no "budget constraint" for this warm glow of giving, it may inflate the value for single studies, and consequently, since the NOAA recommendations (Arrow et al., 1993), it has been suggested that the donation payment vehicle should not be used in the estimation of welfare values. Related to this is also the possible diversion between preferences elicited when being asked to act as 'homo economicus' as opposed to 'homo politicus' (Nyborg, 2000). Nyborg argues that individuals may have a dual set of preferences, one relating to their personal well-being and one related to their subjective social welfare and that the latter may be the one that dominates when answering WTP question where their role as citizens are dominating. If this is the case, the underlying preferences will differ depending on the payment vehicle – as it emphasizes two different roles the individual may take. Another problematic component in the use of tax as a payment vehicle is that tax aversion may influence the choices made by respondents (Kallbekken and Sælen, 2011), including the possibility that a tax being coercive results in crowding-out of intrinsic motivation to contribute to public goods (Goeschl and Perino, 2012). We explicitly test for the effect of payment vehicle by comparing preferences elicited in subsamples where the only difference is the payment vehicle – a hypothetical tax and a hypothetical donation.

Another parameter that is typically varied in comparisons of SP-RP preference studies is the bid range – not least when the revealed comparison is conducted in a lab where individuals are faced with a real budget constraint, whereby participants for ethical reasons cannot be asked to leave the experiment worse off than when they entered. Consequently, any trade-off they are asked to make has to be made with the experimentally provided money (Kagel and Roth, 1995). Previous studies examining the effect of varying the level of the cost attribute in CE studies have typically found higher levels of WTP when the levels of the cost attribute increased (Carlsson and Martinsson, 2008; Ladenburg and Olsen, 2008; Mørkbak et al., 2010), but the literature also contains examples of no difference in WTP (Hanley et al., 2005; Kragt, 2013). Kragt (2013) proposes that the result of no difference in WTP when introducing an increased cost vector could be due to the presence of hypothetical bias, while also mentioning that a high level of the cost attribute might affect the unobserved error variance, leading to higher levels of variance. This effect on error variance has also been proposed by Mørkbak et al. (2010), but Kragt (2013) argues that the effect could also be reversed, in that high costs could lead to more certainty in choice, thus reducing the error variance. As in Kragt (2013), we vary the level of the cost attribute between samples, but the change in bid range in our setting is 10-fold, whereas it is only a 50% increase for the highest bid in her study. Thereby we push the bid range comparison into the extreme.

We keep the survey context constant when investigating the effects of varying elicitation format, payment vehicle and bid range. Respondents are asked to value climate policies, with distributional outcomes manifesting in the far future as well as in present time, across three specific regions of the world. The future distributional outcomes are described in terms of changes to the average income level as a result of lower impacts of climate change resulting from additional investments in climate policy. The present-time outcome is the provision of the co-benefit of cleaner air, brought about by increased climate mitigation activity in the fuel combustion sector. Compared to earlier SP-RP studies, we are able to explicitly distinguish the effect of payment vehicle, and vary the bid range and elicitation format. Furthermore, we are making the comparison for a good which has inherent moral components, which are relevant

[1] Several studies have investigated the performance of incentivized stated preference studies in mitigating hypothetical bias with mixed results. Studies point to factors such as the type of good being valued and whether one uses with-in or between-subjects comparisons of WTP (Carlsson and Martinsson, 2001; Lusk and Schroeder, 2004; Johansson-Stenman and Svedsäter, 2008; Chang et al., 2009).

for many policy-informing SP studies, but also challenging.

2. Method and materials

2.1. Choice context

The data for this study consist of four samples using the same choice context, which is asking respondents to value distributional outcomes of climate policy.[2] Climate policy is described as having a future and a present time impact on outcomes in three specific regions of the world, Western Europe (WE), Southeast Asia (SEA) and Sub-Saharan Africa (SSA), which were chosen as to generate a natural gradient in income. The future outcome of climate policy is defined as occurring in year 2100, a year which is an established reference year in integrated assessment models and in much of the international negotiations on climate change (IPCC, 2014). The future outcome of additional climate policy effort now is described as lowering the expected economic impacts from climate change in 2100, through additional investment in policies that mitigate and adapt to climate change. The baseline scenario (status quo) is that no additional climate policies will be implemented, resulting in an excepted loss in average income of 5% in each of the three regions in 2100. Additional climate policy will lower this expected loss, but never fully remove it. The four levels for the income effect attribute in each region are shown in Table 1 and were set using information from an application of the integrated assessment model FUND (Anthoff and Tol, 2010). The present-time outcome is defined as the provision of a co-benefit from CO_2 mitigation, described qualitatively as “fewer cases of respiratory diseases” in either of three regions.

The respondents were informed that the policies considered would implement CO_2 reduction through changes to fuel combustion technology in the transportation, production and household sector, which as a side-effect to reducing CO_2 also will reduce air pollution, thus generating the co-benefit of fewer people being affected with respiratory diseases resulting from air pollution. The co-benefit attribute could take four levels; an effect in one of the three regions or “no effect”, the latter indicating that the proposed climate policy only included adaptation efforts.

The type of payment vehicle, as well as the bid range of the payment vehicle varied across the different data collections included in the paper. As displayed in Table 1, the payment vehicle could assume a *high bid range*, where the levels ranged from 0 to 2000 DKK, which was the case in two of the three data collections, which used strictly hypothetical payment vehicles; an increase in income tax (not time limited) and a one-time donation. The payment vehicle could also have a *low bid range*, where levels ranged from 0 to 200 DKK, which it did in the last data collection that used a real, one-time donation mechanism as payment vehicle. The real donation mechanism was implemented by endowing respondents with 200 DKK, and then asking them to make a series of choices, of which one would be drawn at random to be realised. The level of 200 DKK was selected considering the financial constraints of the study, while also ensuring a credible payment level for the respondents. The endowment corresponds to an hourly wage of 600 DKK ~80 Euro, which constitutes comparatively high wage when compared to similar studies (Löschel et al., 2013). The respondents were paid the difference between the endowment and the price of the selected option in the realised choice. The money donated to climate policy was then used to purchase and delete CO_2 quotas and credits in the European Emissions Trading System (ETS) and donated to the UN Adaptation Fund.[3] All surveys included budget reminders, as well as informing respondents about the results of the studies being made public, which was done in order to enhance the perceived consequentiality of the respondents' choices (Carson and Groves, 2007).

To ensure comprehensibility of the presented survey, it was pre-tested in two focus groups with laymen with researchers present and two pilot studies (one with students and researchers and one using respondents from the panel that would handle the main survey designs). Besides providing priors for the technical design, the pre-testing improved the representation of the choice context, herein leading to the decision to include aids that supported as many perception preferences as possible. The bid range was pre-tested for both data collections involving one-time donations to ensure that respondents were sensitive to it. Because respondents in the revealed setting were endowed money beforehand, the range is smaller. Yet, we found that respondents reacted to the proposed bid range.[4]

2.2. Data samples and specification of tests

The three data collections included in this paper stem from three separate discrete choice experiments (DCE), which are;

1. **SPT** - Stated Preference Tax, a DCE with a *hypothetical tax* as payment vehicle
2. **SPD** - Stated Preference Donation, a DCE with a *hypothetical donation* as payment vehicle
3. **RPD** for Revealed Preference Donation, a DCE with a *real donation mechanism* as payment vehicle.

[2] Other papers focus on characterizing respondents distributional preferences, which will not be the focus in the present paper. For a more in-depth motivation of the choice context, please refer to these papers (Svenningsen, 2017; Svenningsen and Thorsen, 2017).

[3] Another paper, Svenningsen (2017), focuses explicitly on this data collection and further explains the real donation mechanism and the results stemming from this data.

[4] Figs. 2 and 3 in the Appendix show the distribution of donation choices, from which it is visible that respondents reacted negatively to increases in donation prices.

Table 1
Attributes and levels.

	Levels								Status Quo
Co-benefit	Fewer cases of respiratory diseases (Western Europe)								No effect
	Fewer cases of respiratory diseases (Southeast Asia)								
	Fewer cases of respiratory diseases (Sub-Saharan Africa)								
	No effect								
Income-effect									
- Western Europe	42.000		33.600		16.800		8.400		42.000
- Southeast Asia	21.000		16.800		8.400		4.200		21.000
- Sub-Saharan Africa	10.500		8.400		4.200		2.100		10.500
Price, DKK									
Bid range high (hypothetical tax and donation)	0	100	200	400	600	900	1200	2000	0
Bid range low (real donation)	0	10	20	40	60	90	120	200	0

From these three data collections, four samples are used to construct three separate tests of the effects of elicitation format, bid range and payment vehicle upon the valuation of climate policy. An overview of the link between the data collection, samples and tests can be found in Fig. 1.

In order to examine the effect of different survey designs upon the choice behaviour of respondents, we have specified three different tests that essentially vary which samples are compared to each other. Two of the tests are within-subjects tests, where within refers to the fact that we compare data from the same individual, who is sampled twice under different survey designs. We specify and test population-level models, focusing on the average effects of sampling schemes on WTP, as this is our main interest. To take full advantage of the within-subject comparison, we also discuss individual-specific model comparisons in preference space in Section 3.3. The last test offers a between-subjects test. All three tests are evaluated by two general procedures; 1) the Swait-Louviere procedure which tests for differences in preference structure and scale across samples and 2) willingness-to-pay (WTP) comparisons across samples, herein testing whether the difference in WTP is statistically significant. The four tests are specified as follows:

Test_All: Test_All is our main test, where we compare the two datasets, SPT_1 and RPD. Any differences between models estimated on these two datasets can be attributed to the factors that have been varied; elicitation format, bid range and payment vehicle. This corresponds to a frequent stated-revealed preference comparison in the literature.

Test_PV: For Test_PV, dataset SPT_2 and SPD are compared. The *only* factor that changes between these two datasets is the payment vehicle, which allows a possible difference in the observed models to be attributed to this effect alone. The result from this test will be used as an indicator for the isolated effect of a change in payment vehicle, which then, if taken as given, can be used to infer and exclude the effect of a change in payment vehicle from Test_All, potentially leaving only two factors (elicitation format and bid range) as drivers behind the results of Test_All.

Test_E&B: This test looks at the difference between contributions in the two donation scenarios, SPD and RPD, and offers a between-subject test, as the two samples are not composed of the same individuals. Any differences identified can be attributed to either a change in the elicitation format (going from a hypothetical to real one-time donation) or the bid range, while keeping the payment vehicle constant.

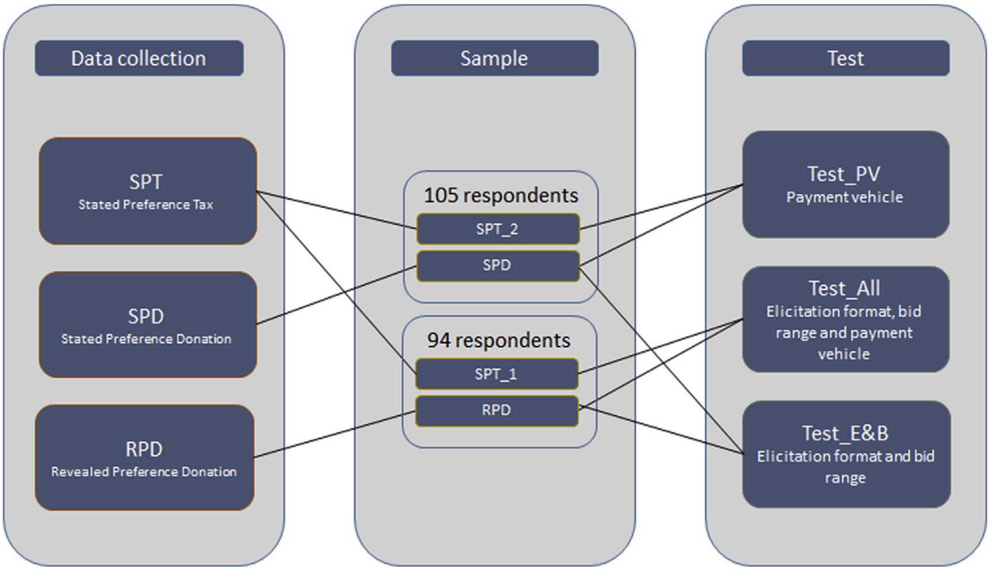

Fig. 1. Overview of data collections, samples and tests included in the paper.

2.3. Technical design and survey designs

All data collections were built around a survey consisting of three sections. The first section introduced the climate policy choice context and elicited general attitudes towards climate change. The second section contained the choice-cards where respondents made a series of climate policy choices. The third section included follow-up questions and elicited socio-demographic information. The experimental design in the SPT and SPD survey designs consisted of 8 choice cards, with three alternatives, distributed into two blocks, leaving a total of 16 unique choice-card designs. The RPD survey consisted of 16 choice-cards again divided into two blocks, leaving a total of 32 unique choice-card designs. The technical designs for all three survey designs were optimized according to D-efficiency using the software Ngene (ChoiceMetrics, 2012). All three designs were built on a main-effect dummy-coded MNL model. The design was identical for SPT and SPD and was built on priors from pilots of SPT. The technical design changed in the RPD survey designs, using updated priors from the SPT data collection.

Data collection was handled by Userneeds,[5] who controls an online panel consisting of more than 95,000 members of the general Danish public. All samples were collected during the period December 2015–March 2016, with approximately two weeks between collection of SPT and SPD, and three weeks between SPD and RPD. Respondents were invited via email. Due to the sampling scheme it is not possible to calculate a standard response rate as respondents were sampled based on quotas representative on parameters such as age, gender and income.[6] The samples included in Test_All and Test_PV are fairly comparable in terms of gender composition and a chi-square test confirms that at a 5% significance level the two samples are not different from each in terms of gender composition. The same applies to the educational levels attained by respondents, for which no significant difference can be found between the samples in Test_All and Test_PV. Notice that as the WTP comparisons in Test_All and Test_PV are based on within-subject samples, it does not affect the results of the test. Test_E&B relies on subsamples from Test_All and Test_PV and here any potentially observed differences in the observed valuation of climate policies could potentially be attributed to the differences in age and income level. Please refer to Tables 8 and 9 in the Appendix for descriptive statistics and chi-square tests.

2.4. Econometric framework

The theoretical framework for modelling respondents' choice data is the Random Utility framework developed by McFadden (1973), which in conjunction with Lancaster's theory of characteristics of demand (Lancaster, 1966) allows us to define the utility of a decision maker n considering choice situation i as consisting of an observable and unobservable part

$$U_{ni} = \beta x_{ni} + \varepsilon_{ni} \tag{1}$$

Here β is a vector of parameter coefficients to be estimated, based on the observed parameters $x_{i,j}$ which can be choice attributes as well as individual characteristics of the decision maker. The unobservable part is the stochastic element ε_{ij}, which is assumed to be type I extreme value distributed, resulting in the definition of the individual choice probabilities over a series of T choices as

$$P_{ni} = \prod_{t=1}^{T} \frac{e^{\beta x_{nit}}}{\sum_j e^{\beta x_{njt}}} \tag{2}$$

This specification results in the Multinomial Logit, where all respondents are assumed to react homogenously to marginal changes in the choice attributes. In the models applied in this paper, we wish to allow for preference heterogeneity and therefore we specify a Random Parameters Logit model,[7] rendering the choice probability over a sequence of T choices to be defined as

$$P_{ni} = \int \prod_{t=1}^{T} \frac{e^{\beta_n x_{nit}}}{\sum_j e^{\beta_n x_{njt}}} f(\beta) d\beta \tag{3}$$

Equation (3) captures that the estimated coefficients β varies over participants, with a density described by $f(\beta)$, and not fixed as in eq. (2) so estimation of this model involves simulation of the log-likelihood (Train, 2009). All attribute parameters are specified as following a normal distribution, including the price parameter. Letting the price parameter follow a normal distribution allows people to react positively to price increases, an assumption that in many applications is highly unreasonable. However, we argue that the choice context presented in this study is strongly focused on moral aspects, where one could expect non-rational responses to increases in price. The observed preferences are likely to contain elements of altruistic behaviour, warm glow or strategic behaviour, and by specifying a distribution for the price parameter that allows for positive values, we allow people to react non-rationally to price increases.[8] In a latent class model based on data from the same experimental context, Svenningsen and Thorsen (2017) finds a non-negligible group of people with a positive price coefficient, a class which they label as the strategic class and the authors discuss the possibility that the presented context in itself might induce some people to signal a commitment to the climate cause.[9] So in order

[5] www.userneeds.dk/markedsanalyse.

[6] Protesters have not been excluded from the data because sensitivity checks showed no changes to the interpretation in the population-level models, although the model fit in terms of LL significantly improved.

[7] As a robustness check we have also run simple MNL models. These are included in the Appendix and they show the same overall interpretation of the parameter coefficients as in the RPL population level models.

[8] It should be noted that we in this study are not able to distinguish between the underlying different motivations.

[9] A reviewer has correctly pointed out that it would interesting to pursue the apparent non-linearity in the preference for the price of climate

to allow for a more accurate reflection of the actual data, in this specific context, a normal distribution was chosen[10]. Choosing a normal distribution for the price parameter has an implication in that in the calculation of WTP based on estimated random parameters, a normally distributed price coefficient will result in undefined moments of the WTP distribution (Hess et al., 2005; Daly et al., 2012), making policy inference about the valuation of attributes problematic. As a consequence of this, we calculate WTP based on median estimates by the Krinsky-Robb procedure.

2.4.1. Sample comparisons

In order to compare results between samples, we conduct the Swait-Louviere test (Swait and Louviere, 1993). It is a two-step procedure with the first step determining whether the assumption of equal preference structure between samples is a reasonable assumption (testing the H1A hypothesis). If an equal preference structure cannot be rejected in step one, then one can proceed to the second step, which tests differences in scale between samples (testing the H1B hypothesis). The test involves estimating two types of models on pooled datasets, where one model explicitly models the relative scale parameter, with the scale for one dataset normalized to 1, while another model essentially assumes no difference in scale and therefore the scale does not enter as a separate parameter in model estimation. The Swait-Louviere test is essentially a Likelihood Ratio test, with a chi-square distributed test value.

The second step used for model comparison is the calculation of the willingness to pay for all six climate policy attributes, and herein also testing whether the WTP estimates are statistically significantly different by conducting a Poe-test (Poe et al., 1994; Poe et al., 2005). The WTP estimates are based on the estimated normally distributed random coefficients, which theoretically has been shown to result in infinite moments for the distribution of WTP (Daly et al., 2012). We therefore follow the suggestion of Bliemer and Rose (2013) and base the WTP comparisons on median WTP, as the median always is finite and therefore can be used as an approximation of the mean WTP.

Lastly, we conduct three robustness checks, which all are discussed in Section 3.3. First, we test whether the conclusions of the Poe test hold with a lognormally distributed price. These results are shown in Table 10 in the Appendix. Second, we estimate models in preference space allowing for individual specific estimation of hypothetical bias of each attribute. These results are shown in Table 12 in the Appendix. Lastly, we investigate potential learning effects arising from the respondents being resamples across the different data collections.

All models are estimated in Stata (StataCorp, 2013), using the *mixlogit* (Hole, 2007) and *gmnl* command (Gu et al., 2013), each simulated through Halton draws, using 1000 replications.

3. Results

We start by showing the model estimates. Then we conduct the Swait & Louviere test of heterogeneity in preference structure and scale for all three specified tests and as a final test, we estimate and compare WTP of the different samples included in each of the three tests.

Table 2 lists the results of the four models. In all models, the estimated attribute parameters have the expected signs. The cost coefficients are all estimated as negative, indicating that respondents on average experienced disutility from increases in the price of climate policy. The relative cost coefficient is notably higher in the incentivized stated preference sample, RPD, suggesting a stronger reaction to price when respondents were making their choice of policy based on an actual endowment of money. Across all four models, future income effects in all regions, WE, SEA and SSA generate disutility, indicating that respondents dislike policies that generate income losses for people living in these regions in the future. In all models, except SPT_1, this effect is significant for all regions at a 5% significance level. The provision of co-benefits in all three regions suggests that respondents on average gain utility from policies that provide this benefit in all regions and the effect is significant in all regions at a 5% level in all models except SPD. The utility effect of the status quo alternative of no additional climate policy (asc) generate disutility in all four models for the respondents, with the effect being statistically significant at 5% in all models except the samples based on incentivized stated preference elicitation (RPD).

Table 2 also suggests the presence of significant preference heterogeneity for many of the random parameters, with variation across the four estimated models. The size of the preference heterogeneity relative to the mean coefficient is stable across models for income effects in WE (incWE) and the price of the policy (price). All models indicate that preferences for co-benefits provided in SSA (cobSSA) are homogenous.

(*footnote continued*)
policy, and we have run additional models that discretize the price parameter, in order to examine the reasonableness of assuming that some group of respondents do exhibit a positive preference for increases in price. These models are available from the author by request. The models confirm the presence of a group of respondents for whom increases in the price of climate policy has a positive impact. This positive preference can be thought to be driven by altruism, warm glow or strategic behaviour, but unfortunately we are not able to precisely identify the underlying driver. We therefore make reference to all motivational factors throughout the paper.

[10] The model-testing phase further showed small differences in model fit (log-likelihood) compared to models with a lognormal or triangular distribution for the price parameter.

Table 2
Random parameters logit models with full covariance matrix, estimated for the four samples. STD. Errors in parentheses. Significance level: *P < 0.05, ** P < 0.01, ***P < 0.001.

	SPT_1	RPD	SPT_2	SPD
Mean				
asc	−4.850**	−0.886	−5.400**	−3.273***
	(1.592)	(0.735)	(1.924)	(0.815)
incWE	−0.041***	−0.039***	−0.025***	−0.030**
	(0.008)	(0.008)	(0.007)	(0.010)
incSEA	−0.021	−0.049***	−0.028**	−0.059***
	(0.011)	(0.011)	(0.010)	(0.013)
incSSA	−0.058*	−0.072***	−0.047*	−0.074**
	(0.024)	(0.020)	(0.020)	(0.025)
cobWE	1.878***	1.472***	1.287***	1.471***
	(0.283)	(0.233)	(0.195)	(0.281)
cobSEA	0.965***	1.185***	0.664***	0.340
	(0.189)	(0.176)	(0.156)	(0.182)
cobSSA	1.144***	1.438***	0.633***	0.171
	(0.233)	(0.184)	(0.192)	(0.225)
price	−1.239***	−20.841***	−0.603**	−1.253***
	(0.324)	(4.617)	(0.186)	(0.314)
Standard Deviation				
asc	5.382***	4.308***	6.596***	4.119***
	(1.154)	(0.572)	(1.473)	(0.874)
incWE	−0.031*	0.052***	0.032***	0.052***
	(0.013)	(0.008)	(0.010)	(0.014)
incSEA	−0.007	0.053**	−0.000	0.041
	(0.059)	(0.017)	(0.029)	(0.033)
incSSA	0.081	0.058	0.066	0.065
	(0.060)	(0.036)	(0.050)	(0.068)
cobWE	−1.116**	1.166***	0.052	1.326***
	(0.388)	(0.241)	(0.967)	(0.374)
cobSEA	−0.004	0.583**	0.029	0.059
	(0.274)	(0.211)	(0.380)	(0.894)
cobSSA	0.022	0.457	0.016	0.077
	(0.412)	(0.286)	(0.283)	(0.397)
price	2.171***	31.823***	1.238***	2.337***
	(0.355)	(5.732)	(0.225)	(0.350)
Choice obs	2256	4512	2520	2520
BIC	1080.034	1929.308	1282.316	1242.638
LL	−478.246	−897.338	−578.502	−558.663
Pseduo R2	0.27	0.38	0.26	0.26

3.1. Swait & Louviere tests

To investigate whether the overall preference structure and scale differed between samples, a series of Swait-Louviere tests were conducted. The log-likelihood and test values included in the Swait-Louviere test for all four test-specifications can be found in Table 3. Since we are in effect comparing answers from the same individual across different sampling schemes, any variation in scale can with some reason be said to be a result of differences in survey design.

The H1A test value in Test_All is 137.212 which with a critical χ^2_{16} value of 23.6 allows us to reject the null-hypothesis of equal

Table 3
The table displays the log-likelihood from both individual models included in each test (LL$_1$ – LL$_3$) and from the pooled models (LL pooled). The chi-square values from the Swait Louivere test (H1A and H1B) is also displayed.

	Within-subject		Between-subject
	Test_All	Test_PV	Test_E&B
LL pooled - scale modelled	−1444.190	−1136.717	−1547.919
LL$_1$	−478.246	−578.502	−897.338
LL$_2$	−897.338	−558.663	−558.663
LL$_3$			
H1A: chi –square value	137.212***	−0.896	183.836***
LL pooled - scale not modelled	−1462.027	−1148.104	−1563.288
H1B: chi square value		22.774***	

preference structure between SPT_1 and RPD, suggesting that the different survey designs significantly influenced the modelled preference structure. The rejection of equal parameter structure is also confirmed for Test_E&B. For Test_All and Test_E&B, the difference in sampling scheme involves a change in elicitation format, going from stated to an incentivized stated preference setting, which the results from our models indicate causes significant differences in the estimated preference structures. Consequently, a common scale difference between them cannot be estimated.

The isolated effect of changing a (hypothetical) payment vehicle is tested in Test_PV, in which the H1A of equal parameter structure between sample SPT_2 and SPD cannot be rejected. However, the assumption of equality of scale between these two samples can be rejected (H1B chi-square value 22.77 against a critical χ_1^2 value of 3.84), with the error variance in the SPT_2 sample being higher than in the SPD sample. This indicates that the payment vehicle has a significant influence on the observed error variance, with respondents making more certain choices when they are asked to value policy alternatives that use a donation as payment vehicle, compared to policies that use tax payment as payment vehicle.[11]

3.2. WTP comparisons between samples

To examine how differences in survey design influenced respondents' valuation of the climate policy attributes, Table 4 compares the median marginal WTP of each attribute across the four models. The confidence intervals are calculated using the Krinsky Robb procedure with 800 repetitions. Using the Krinsky Robb procedure ensures that percentiles are defined, in the case the moments of the WTP distributions are not defined (Bliemer and Rose, 2013). Starting with Test_All, Table 4 indicates a higher negative WTP for income effects in all three regions, as well as a higher positive WTP for the provision of co-benefits in all three regions in the hypothetical taxation context compared to the real donation context. For Test_PV the results indicate a more comparable WTP for income effects in all three regions, whereas the WTP for co-benefits in all regions is lower in the hypothetical donation context, compared to the hypothetical taxation context. Test_E&B is a between-subjects comparison of WTP, relying on results from the models RPD and SPD in Table 4. Comparing the WTP from these models indicates that the WTP is higher in a hypothetical donation context (SPD) compared to a real donation context (RPD).

To test whether the estimated distributions of WTP are statistically significantly different between samples, Table 5 presents the results of a series of tests on the equality of the estimated WTP's from the different samples using the convolution approach (Poe et al., 1994; Poe et al., 2005). The following conclusions are based on a 10% significance level. In Test_All, the results of the Poe tests indicate significant differences in WTP between SPT_1 and RPD for all six policy attributes, indicating that the WTP in a hypothetical elicitation context, using tax as payment vehicle with high bid range, generates significantly higher valuation of all policy attributes, compared to an incentivized elicitation context, with a real donation as payment vehicle, using a lower bid range.

Looking at Test_PV which isolates the effect of a change in payment vehicle, the results indicate that the WTP for future income effects is not significantly different across the two sampling schemes, whereas the valuation of the more immediate outcomes of climate policy, the provision of co-benefits, is significantly higher in the hypothetical taxation context. For Test_E&B, the results of Table 5 suggest a significantly higher valuation of all climate policy attributes in a hypothetical donation context, compared to a real donation context.

To sum up, the results when comparing and testing difference in WTP across samples suggest significant differences in WTP in several of the proposed tests. The results indicate that a change in payment vehicle (Test_PV) does not influence the valuation of a future moral good (income effects), whereas a change in payment vehicle influences the valuation of a more immediate moral good (provision of co-benefits). Taking the results of Test_PV as a given, this then suggests that the observed difference in Test_All across all attributes is caused by either a change in elicitation format (stated vs. incentivized stated) or by a change in bid range. Furthermore, the results of a between-subjects test of changes in elicitation format and bid range with the payment vehicle kept constant (Test_E&B) also indicates that the valuation of climate policy attributes is significantly higher in a hypothetical context, compared to a real donation counterpart. Thus we conclude that, in the context of moral goods, we find evidence of a hypothetical bias in the form of significant differences between our samples, and that these differences are likely driven by differences in the bid range or by the elicitation format – i.e. whether it is hypothetical or not. Specifically, the differences are more pronounced for future moral goods than for immediate moral goods.

3.3. Robustness checks

3.3.1. Lognormal price

The assumption of the price parameter following a normal distribution is arguably most influential in the above WTP comparisons, and although care has been taken to avoid the pitfalls of an unidentified distribution of WTP by comparing the median rather than the mean, it is nevertheless of interest to examine whether the conclusions hold under a different distributional assumption. Table 10 in the Appendix list the results from Poe tests, where the price parameter is assumed to be log-normally distributed, a distributional form proposed by Daly et al. (2012) that avoids undefined moments for the distribution of WTP. The results of Table 10 indicate no significant difference for the conclusions reached in Test_All and Test_E&B, but for Test_PV we observe an interesting change in interpretation, with the results suggesting a statistically significantly higher WTP for all six attributes in the hypothetical

[11] This difference in error variance could of course also be attributed to a learning effect, because respondents in SPD have participated in a similar study before (SPT_2), and therefore are more certain of their choices. This effect is examined in Section 3.3.3.

Table 4
Median WTP and confidence interval, DKK for all four models. 95% confidence interval in brackets, calculated using the Krinsky Robb procedure. Test_E&B is a comparison of RPD and SPD, and is therefore excluded from the Table.

Median WTP	Test_All		Test_PV	
	SPT_1	RPD	SPT_2	SPD
incWE	−0.0338 (0.0007 to −0.0684)	−0.0019 (−0.0003 to −0.0035)	−0.0435 (0.0259 to −0.1128)	−0.0245 (0.0059 to −0.0549)
incSEA	−0.0169 (0.0126 to −0.0464)	−0.0024 (−0.0002 to −0.0045)	−0.0470 (0.0294 to −0.1233)	−0.0473 (−0.0006 to −0.0941)
incSSA	−0.0459 (0.0278 to −0.1197)	−0.0034 (0.0001 to −0.0069)	−0.0749 (0.0792 to −0.2290)	−0.0589 (0.0218 to −0.1396)
cobWE	1537.7 (2.8439–0.2315)	71.8 (0.1340–0.0096)	2167.5 (4.8435 to −0.5084)	1175.9 (2.2817–0.0702)
cobSEA	788.6 (1.6162 to −0.0390)	57.2 (0.1020–0.0123)	1134.7 (2.9155 to −0.6461)	269.1 (0.7470 to −0.2088)
cobSSA	947.3 (1.9050 to −0.0104)	68.7 (0.1272–0.0103)	1094.9 (2.7639 to −0.5742)	136.8 (0.6646 to −0.3911)

Table 5
Results of Poe test for the three specified test. The P-VALUE refers to whether the WTP in the first-mentioned sample in row two, is significantly larger. As an example, in Test_All, the P-value refers to the WTP of the SPT sample being significantly larger than the RPD sample for all six estimated climate policy attributes.

	Test_All	Test_PV	Test_E&B
	SPT-RPD	SPT-SPD	SPD-RPD
	P-value for diff.	P-value for diff.	P-value for diff.
incWE	0.000	0.164	0.000
incSEA	0.058	0.507	0.000
incSSA	0.012	0.371	0.001
cobWE	0.000	0.078	0.000
cobSEA	0.000	0.012	0.001
cobSSA	0.000	0.009	0.015

donation context, compared to the hypothetical taxation context. With a normal distributed price, Test_PV indicated that for income effects there was no statistically significant difference between a hypothetical taxation and donation context, whereas the test indicated that the WTP for co-benefits in all three regions were *higher* in the hypothetical taxation context. As previously discussed, the change in payment vehicle could produce opposite effects on WTP, e.g. a non-coercive context such as the hypothetical donation could both induce free-riding and implicitly lower WTP compared to the coercive taxation counterpart, but one could also expect that respondents were more susceptible to warm glow effects in the donation context, thus having a higher WTP than when asked about a hypothetical tax increase. If the latter effect is dominating, restricting the distribution of the price parameter to only assume negative values, essentially forces every respondent with a positive coefficient on the price close to zero in the lognormal distribution, thus producing large WTP estimates. We propose that mechanism could explain why we observe this apparent change in WTP between the two hypothetical contexts, which highlights that although the assumption of a normally distributed price parameter has its theoretical downsides, the choice context itself contains important information regarding what could be assumed about how people react to price. We argue that the present choice context is strongly focused on morality, which could cause people to react differently to increases in price than compared to other contexts such as transportation choices.

3.3.2. *Individual level effects models*

To take advantage of the within-subject comparisons, we also estimated individual specific models, where the utility is described as in the models in Table 2, but adding interaction terms between elicitation format and all climate policy attributes, thereby directly testing in preference space how the individual valuation of the attributes differ between survey designs. If the interaction term is significant, the preference for the given parameter is significantly different in the two samples. Results are shown in Table 12 in the Appendix and largely confirm the findings discussed above. Model SPT_RPD pools the data from the same respondents participating in a stated preference study, using taxation with high bid range as payment vehicle with data from the real donation context using low bid range as payment vehicle. In SPT_RPD we find a statistically significant effect of individuals valuation of the baseline alternative (interaction term “asc_hyptax”), suggesting that individuals in a hypothetical context experience significantly more disutility of not supporting climate policies. The results furthermore indicate that individuals in a hypothetical context are less price-sensitive (interaction term “p_ hyptax”), compared to a situation where they are making decisions with real monetary consequences. Interestingly, the individual specific models indicate that the hypothetical bias only manifests on attributes in the respondent's own region. The results indicate that respondents in a hypothetical context experienced significantly more disutility from future income losses in their own region (interaction term “iwe_ hyptax”), while also benefitting significantly more from the provision of co-benefits in their own region (interaction term “cwe_ hyptax”). Model SPT_SPD captures the isolated effect of a change in payment vehicle and the individual specific models confirm the result of the Swait-Louviere test, as we find no statistically significant interaction terms in this model, indicating no difference in the individual valuation of attributes.

Table 6
Swait-Louivere tests of learning effects.

	Between-subjects	
	SPD (resample – new-sampled)	RPD (resample – new-sampled)
LL pooled: scale modelled	−1133.220	−1982.172
LL_1	−558.663	−897.338
LL_2	−578.021	−962.035
H1A: chi –square value	−6.928	245.598
LL pooled: scale not modelled	−1148.043	−1984.344
H1B: chi square value	29.646	

3.3.3. Learning effects

An underlying question regarding the results presented so far is whether the observed differences in preference structure, scale and attribute valuation, is caused by respondents being resampled, leaving their choices to be influenced by learning effects. Learning effects have been a topic of interest in the stated preference literature for many years and refer to the finding that respondents in a series of valuation questions learn about the good they are valuing, herein also their preferences for the good (Bateman et al., 2008). The typical application under which learning effects have been investigated is through examining the choices of respondents *within* a survey (Holmes and Boyle, 2005; Czajkowski et al., 2014), while we wish to apply the term to capture any differences in observed behaviour of the same respondent *between* surveys compared to respondents only participating in one survey. The between survey focus has previously been used to explore the individual stability of preferences over time (Mørkbak and Olsen, 2015) and while the stability of preferences is a relevant topic, we are more interested in discovering whether familiarity with an abstract good influence individual valuation. The expectation is that having experience with valuing a somewhat abstract good as the one considered in this survey will result in different behaviour compared to individuals who have no experience with the good being valued.

The data sampling scheme included subsamples of resampled (respondents who had participated in the SPT data collection) *and* new-sampled respondents (respondents who had not participated in a previous version of the survey) in both the SPD and RPD data collections. To test whether having participated in an earlier version of the survey could be driving the results, we condition a test of learning effects. The data sampling scheme allows us to test learning effects in both a hypothetical and an actual donation context, thereby examining the link between learning effects and sampling schemes. For the test of learning effects in a hypothetical donation context (SPD), we compare the choices of 105 resampled respondents, to the choices of 106 new-sampled respondents. For the test of learning effects in an actual donation context (RPD), we compare the choices of 94 resampled respondents to the choices of 101 new-sampled respondents. Table 6 contains the results of the Swait-Louviere tests and Table 7 the Poe tests on the difference in WTP between resampled and new-sampled respondents in the SPD and RPD data collections.

The results of the Swait-Louviere test indicate that the preference structure is significantly different between resampled and new-sampled respondents in the revealed donation context (RDP). Interestingly, the results do not indicate a difference in preference structure between resampled and new-sampled respondents in the hypothetical donation context (SPD), suggesting that the linkage between the sampling scheme and learning effects might not be so straightforward.

The results do indicate a significant difference in scale between resampled and new-sampled respondents in the hypothetical donation context (SPD), indicating that the error variance was actually larger for resampled respondents compared to new-sampled respondents. From an intuitive perspective, this result is puzzling as we would expect, based on learning effects, that respondents,

Table 7
Poe tests - learning effects.

	Between-subjects	
	SPD (resample – new-sampled)	RPD (resample – new-sampled)
	P-value for diff.	P-value for diff.
incWE	0.484	0.159
incSEA	0.478	0.664
incSSA	0.464	0.358
cobWE	0.515	0.188
cobSEA	0.515	0.326
cobSSA	0.517	0.195

*Notes: P-value refers to the test of the first sample in row two being larger than the second sample.

who have been subjected to the context before, are more certain of their choices. On the other hand, having participated in an almost identical survey before with the payment vehicle changing, the resampled respondents from the SPD sample might be more prone to confusion, given that they now are asked to consider a different payment vehicle for the same set of defined policy attributes, whereas the new-sampled respondents only have considered implementing the presented policies through a hypothetical donation. The results indicate that learning effects are sensitive to other data sampling factors than simply having participated in a similar study before. In particular, the results indicate that changes in payment vehicle, in a purely hypothetical context, influence the scale factor and not the overall preference structure.

The observed difference in preference structure for resamples and new-sampled respondents in the revealed donation context does not emerge when testing whether the distribution of WTP between the two samples is different. The Poe test does not indicate a significant difference in WTP for any of the six attributes across resampled and new-sampled respondents in neither the SPD or RPD contexts, suggesting that at least in the current context and design, learning effects do not influence the valuation of the individual policy attributes.

4. Concluding discussion

The presence of hypothetical bias has been an issue since the use of stated preference elicitation methods started. While in some fields, e.g. transportation, validation by the use of revealed preference methods is possible, this is a less viable option for many of the environmental goods where non-use values play a large role. Further, many environmental goods involve asking respondents to act as what Nyborg (2000) calls “Homo politicus” rather than “Homo economicus”, thereby potentially having a dual set of preferences. Stated and revealed preference comparisons in these contexts require explicitly accounting for any differences between elicitation methods, in order to isolate the possible causes of hypothetical bias. In this study we have conducted systematic comparisons of preferences for climate policies, varying both the elicitation format (what could be called the sole hypothetical bias), bid range and payment vehicle, while keeping the choice context constant.

We isolate the effect of payment vehicle and find that in most cases, the difference in payment vehicle is not causing differences in WTP between samples. Especially when the good in question relates to distant benefits (income effects in the future), this is the case. This may be a bit surprising in that one could expect the “Homo politicus” set of preferences to be more pronounced when a tax is used as a payment vehicle. On the other hand, there are also diverging factors affecting this difference as discussed previously, e.g. the warm glow of giving which is especially present in the revealed donation context and drives WTP up, free-riding behaviour in the donation context which acts in pushing revealed WTP down, and tax aversion pushing the hypothetical WTP down. Furthermore, although an income tax as suggested here as a payment vehicle for climate policies is a well-known policy instrument in Denmark to acquire the needed funding, there is also an ongoing debate on the use of environmental taxes which also affects behaviour.

At the same time in relation to the donation payment vehicle, Denmark has seen a public debate about the effect of the annulment of CO2 quotas in the European Carbon Trading Scheme, with opponents heavily questioning the instrument and arguing that the climate impact is negligible (Silbye and Sørensen, 2017). So while the mechanisms of both payment vehicles are in fact in place and real, people may have seen them as somewhat inconsequential or at least uncertain, probably also because the benefits first come far into the future. Yet, we did not identify this in the pretesting of the questionnaire. From a policy perspective, the negligible effect of a change in payment vehicle it is however reassuring – our results suggests that it is the policy effect rather than the mechanism for achieving this effect that is being valued.

Because of little or no effect of payment vehicle, differences in WTP between samples must either be due to the bid range or the elicitation format. Unfortunately, we are not able to separate the individual effect of these two factors, due to concerns that designing a hypothetical taxation setup with a very low bid range would not allow us to close the demand curve (i.e. too many would say yes to the highest bid), and the research budget did not allow for actual donations using the high bid range. Consequently, we conclude that in our setting, we do find a difference between WTP in a purely stated context compared to an incentivized stated context, indicating the presence of hypothetical bias, and we argue that this is mainly due to either differences in elicitation format or the bid range. Furthermore, our results show differences in the degree of hypothetical bias depending on the characteristics of the good being valued, and consequently it is not possible to derive a general estimate of the bias. This calls for hypothetical bias tests based on the type of good being valued, a result that corroborates previous findings in the literature.

Our study furthermore addresses the issue of using a donation as payment vehicle. Our results indicate that we have to acknowledge that using a donation as an instrument in itself may induce warm glow of giving, causing the usual assumption of a negative price coefficient to be questionable. This is important for policy conclusions, as robustness checks have shown that assuming price to be lognormally as opposed to normally distributed, results in a reversal of the conclusions reached regarding the influence of a change in payment vehicle upon hypothetical bias. Forcing a negative price coefficient as done by the use of a lognormal distribution may lead to exaggeration of differences when one of the samples involves a donation with inherent moral components.

We argue that warm glow, altruistic or strategic behaviour are likely candidates for explaining non-rational reactions to increases in price, in the context of moral goods. In the specification of our econometric model, we have allowed these drivers to be present, by specifying a normal distribution for the price parameter. As discussed, the assumption of normality affects the results as seen from the robustness check using a lognormal distribution. However, because a discrete modelling of price does show a positive price parameter, we would argue that a lognormal distribution in itself could also drive our results and therefore, in this specific context, a normally distributed price parameter is more appropriate.

Some caveats should also be mentioned. The choice context is arguably abstract and complex for respondents which challenge the generalizability of our results into all domains where stated preference methods typically are employed to estimate welfare changes. However, as previously argued, many environmental goods hold a strong moral component, like the good presented to respondents in this study. Furthermore, the two main payment vehicles compared, a hypothetical tax and a real donation, differ in their timing. The hypothetical tax was specified as being annual with no ending point, which corresponds to how most taxes are designed and aligned with the requirements of a proposed policy, whereas the donation was implemented as a one-time payment. Given that the levels of the income effects and provision of co-benefits remained constant across this change in timing of payment, the scenario description in the real donation context is arguably unrealistic, as this scenario lets individuals secure the same policy outcomes at a much lower price. If respondents have realised this, then their WTP in the real donation context should perhaps rather be interpreted as signalling support for a moral good. This could also be a general interpretation of the expressed WTP throughout the study, which was focused on the relative weighting of distributional outcomes across different world regions. There is no doubt that the policy context explored in this paper is more abstract and distant than in most other surveys, given the focus on investigating income effects from climate policies displaced in both time and space. This may drive some of the observed differences between the hypothetical and incentivized studies, causing the observed difference/hypothetical bias to be larger than in contexts involving more familiar goods. However, we find that differences between samples are more pronounced for the present-time provision of co-benefits compared to the distant, future income effects, which suggest that this potential driver is only a minor issue.

Finally, as some of our respondents were resampled, there is a possibility that results could be affected by learning effects. Unfortunately, our sampling scheme did not include randomization of the order in which resampled respondents participated in the different versions of the survey. This would have been ideal as potential order effects could have been identified. However, as two of the data collections included resampled and new-sampled respondents, this paper has also included a test for potential learning effects of having been resampled in a previous version of the survey. The results indicate that learning is present, but that it does not drive the results found.

Hypothetical bias remains a challenge for welfare estimates and the range and interaction between design features that influence hypothetical bias is complex. Our results contribute to the literature on hypothetical bias, by providing a systematic analysis that, within the same survey context, tests several different design factors influence on hypothetical bias. They indicate that the use of WTP estimates, regardless of whether these are based on stated or incentivized stated preference methods, is problematic in the context of strong moral goods. It also shows that researchers should be careful in designing experimental protocols that keep payment vehicle and bid range identical across survey design if the goal is to identify the size of hypothetical bias. Hypothetical bias is likely to depend on the type of goods being valued. While it may be relatively simple to compare it for private consumer goods, we suggest that it is investigated further for goods with a strong moral context as here, but also for issues of high policy relevance and large non-use values like conservation. Such approaches can be both systematic experiments varying one factor at a time, or qualitative studies looking into the motivation behind WTP statements.

Acknowledgements

The authors would like to thank Julian Sagebiel for providing his user-written Stata command, *poetest*. Furthermore, we thank three anonymous reviewers for constructive comments. Jette Bredahl Jacobsen acknowledge the funding from the Danish National Research Foundation for the Centre for Macro-ecology, Evolution and Climate (grant number DNRF96).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jocm.2018.08.001.

APPENDIX

Table 8
Descriptive statistics of samples – sample averages.

	Test_All SPT_1 and RPD	Test_PV SPT_2 and SPD	Population of Denmark
Female	0.49	0.46	0.50
Age	47.39	47.08	41.1
Income[a]	250,000–274,999	250,000–274,999	261,323
Education - Tertiary	0.23	0.22	0.27
Education - Secondary	0.11	0.11	0.09
Education – vocational	0.61	0.60	0.30
Education – primary	0.04	0.06	0.27

Table 9
Descriptive statistics of samples - sample counts and chi-square values.

	Test_All SPT_1 and RPD	Test_PV SPT_2 and SPD	Chi-square test values
N	94	105	
Female	46	48	
Male	48	57	5.082
Age: 18–35	20	32	
Age: 36–56	42	33	
Age: 57–70	30	38	454.399
Income: 0–74,999	0	2	
Income: 75,000–174,999	20	19	
Income: 175,000–299,999	22	39	
Income: 300,000–499,999	43	41	
Income: > 500,000	5	2	229.929
Education: Tertiary	22	23	
Education: Secondary	10	12	
Education: Vocational	57	63	
Education: Primary	4	6	4.267

Note: Some respondents did not want to disclose their income or did not remember and these are not included in the above counts for the income variable. Likewise some respondents chose to indicate their educational level as “other”, and these respondents are not included in the above counts.

Table 10
P-values for poe tests, lognormal price.

	Test_All	Test_PV	Test_E&B
	SPT-RPD	SPD-SPT	SPD-RPD
	P-value for diff.	P-value for diff.	P-value for diff.
incWE	0.000	0.001	0.000
incSEA	0.018	0.001	0.000
incSSA	0.016	0.004	0.002
cobWE	0.000	0.000	0.000
cobSEA	0.000	0.008	0.004
cobSSA	0.000	0.041	0.034

*Notes: P-value refers to the test of the first sample in row two being larger than the second sample. For example in Test_All the p-value refers to the probability that the WTP in SPT is larger than in RPD.

Table 11
Multinomial logit models, STD. errors in parentheses. significance level: *P < 0.05, ** P < 0.01, ***P < 0.001.

	SPT_1	RPD	SPT_2	SPD
asc	−0.154	0.860***	−0.072	−0.324
	(0.188)	(0.145)	(0.172)	(0.171)
income_WE	−0.024***	−0.025***	−0.020***	−0.017***
	(0.004)	(0.003)	(0.004)	(0.004)
income_SEA	−0.014	−0.030***	−0.019*	−0.028***
	(0.008)	(0.006)	(0.007)	(0.007)
income_SSA	−0.040**	−0.041***	−0.034*	−0.049***
	(0.015)	(0.012)	(0.014)	(0.015)
coB_WE	1.356***	1.039***	1.003***	1.031***
	(0.173)	(0.116)	(0.151)	(0.152)
coB_SEA	0.789***	0.743***	0.585***	0.380**

(continued on next page)

Table 11 (*continued*)

	SPT_1	RPD	SPT_2	SPD
	(0.145)	(0.110)	(0.132)	(0.130)
coB_SSA	0.893***	0.993***	0.524***	0.332*
	(0.173)	(0.116)	(0.156)	(0.157)
price	−0.444***	−5.035***	−0.231**	−0.369***
	(0.082)	(0.594)	(0.076)	(0.076)
N	2256	4512	2520	2520
BIC	1372.112	2984.583	1625.242	1579.051
LL	−655.171	−1458.634	−781.293	−758.197

Table 12
Individual specific models. models are built on pooled datasets from test 1–3.

		SPT_RPD	SPT_SPD
Mean			
asc		−1.200*	−2.974***
		(0.496)	(0.648)
income_WE		−0.029***	−0.031***
		(0.005)	(0.007)
income_SEA		−0.036***	−0.047***
		(0.008)	(0.010)
income_SSA		−0.050***	−0.062**
		(0.014)	(0.021)
coB_WE		1.235***	1.247***
		(0.156)	(0.206)
coB_SEA		0.897***	0.388*
		(0.126)	(0.158)
coB_SSA		1.152***	0.268
		(0.137)	(0.194)
price		−7.327***	−0.875***
		(0.890)	(0.199)
asc_hyptax		−1.900***	0.197
		(0.349)	(0.337)
iwe_ hyptax		−0.019*	0.003
		(0.008)	(0.008)
isea_ hyptax		0.015	0.017
		(0.013)	(0.013)
issa_ hyptax		−0.010	0.018
		(0.026)	(0.027)
cwe_ hyptax		0.784**	0.049
		(0.282)	(0.271)
csea_ hyptax		0.219	0.268
		(0.228)	(0.219)
cssa_ hyptax		0.039	0.276
		(0.270)	(0.271)
p_ hyptax		4.989***	0.003
		(0.777)	(0.152)
Standard Deviation			
Asc	4.226***		3.600***
	(0.546)		(0.499)
income_WE	0.030***		0.033***
	(0.004)		(0.007)
income_SEA	−0.032***		0.032*
	(0.009)		(0.014)

(*continued on next page*)

Table 12 (*continued*)

	SPT_RPD	SPT_SPD
income_SSA	−0.039	0.076**
	(0.022)	(0.025)
coB_WE	0.721***	−0.665***
	(0.134)	(0.199)
coB_SEA	−0.174	−0.254
	(0.217)	(0.265)
coB_SSA	−0.264	−0.044
	(0.177)	(0.289)
price	4.102***	1.607***
	(0.613)	(0.205)
N	6768	5040
BIC	3303.374	2483.799
LL	−1545.847	−1139.598

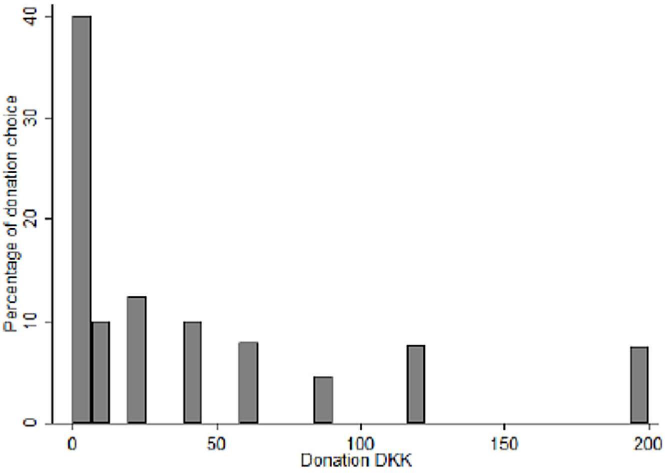

Fig. 2. Showing the distribution of money (DKK) donated to climate policy in the RPD sample.

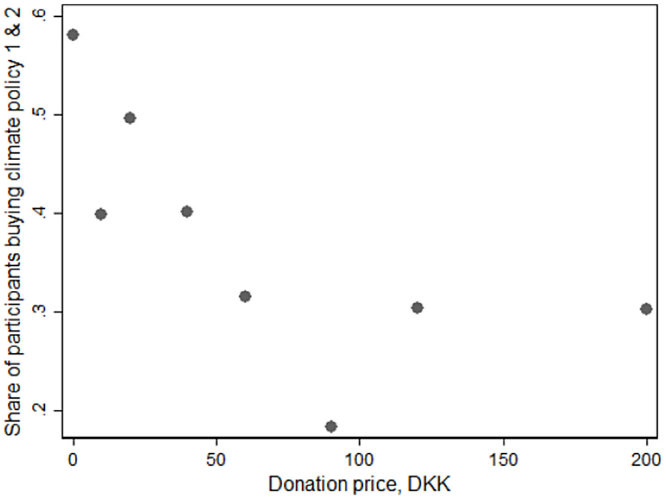

Fig. 3. The choice of climate policy as a function of the donation price. The vertical axis measures the share of pro-climate policy choices at each price level, in the RPD sample.

References

Alpizar, F., et al., 2008. Does context matter more for hypothetical than for actual contributions? Evidence from a natural field experiment. Exp. Econ. 11 (3), 299–314.

Andreoni, J., 1990. Impure altruism and donations to public goods: a theory of warm-glow giving. Econ. J. 100 (401), 464–477.

Anthoff, D., Tol, R.S.J., 2010. On international equity weights and national decision making on climate change. J. Environ. Econ. Manag. 60 (1), 14–20.

Arrow, K., et al., 1993. Report of the NOAA panel on contingent valuation. Fed. Regist. 58 (10), 4601–4614.

Bateman, I.J., et al., 2008. Learning design contingent valuation (LDCV): NOAA guidelines, preference learning and coherent arbitrariness. J. Environ. Econ. Manag. 55 (2), 127–141.

Bliemer, M.C.J., Rose, J.M., 2013. Confidence intervals of willingness-to-pay for random coefficient logit models. Transp. Res. Part B Methodol. 58 (Suppl. C), 199–214.

Carlsson, F., Martinsson, P., 2001. Do hypothetical and actual marginal willingness to pay differ in choice experiments? J. Environ. Econ. Manag. 41 (2), 179–192.

Carlsson, F., Martinsson, P., 2008. How much is too much? Environ. Resour. Econ. 40 (2), 165–176.

Carson, R., Groves, T., 2007. Incentive and informational properties of preference questions. Environ. Resour. Econ. 37 (1), 181–210.

Carson, R.T., et al., 1996. Contingent valuation and revealed preference methodologies: comparing the estimates for quasi-public goods. Land Econ. 72 (1), 80–99.

Carson, R.T., Mitchell, R.C., 1995. Sequencing and nesting in contingent valuation surveys. J. Environ. Econ. Manag. 28 (2), 155–173.

Chang, J.B., et al., 2009. How closely do hypothetical surveys and laboratory experiments predict field behavior? Am. J. Agric. Econ. 91 (2), 518–534.

ChoiceMetrics, 2012. Ngene 1.1.1 User Manual & Reference Guide. Australia.

Crowne, D.P., Marlowe, D., 1960. A new scale of social desirability independent of psychopathology. J. Consult. Psychol. 24 (4), 349–354.

Czajkowski, M., et al., 2014. Learning and fatigue effects revisited: investigating the effects of accounting for unobservable preference and scale heterogeneity. Land Econ. 90 (2), 324–351.

Daly, A., et al., 2012. Assuring finite moments for willingness to pay in random coefficient models. Transportation 39 (1), 19–31.

Epley, N., Dunning, D., 2000. Feeling "holier than thou": are self-serving assessments produced by errors in self- or social prediction? J. Pers. Soc. Psychol. 79 (6), 861–875.

Goeschl, T., Perino, G., 2012. Instrument choice and motivation: evidence from a climate change experiment. Environ. Resour. Econ. 52 (2), 195–212.

Gu, Y., et al., 2013. Fitting the generalized multinomial logit model in Stata. STATA J. 13 (2), 382–397.

Hanley, N., et al., 2005. Price vector effects in choice experiments: an empirical test. Resour. Energy Econ. 27 (3), 227–234.

Hensher, D.A., 2010. Hypothetical bias, choice experiments and willingness to pay. Transp. Res. Part B Methodol. 44 (6), 735–752.

Hess, S., et al., 2005. Estimation of value of travel-time savings using mixed logit models. Transport. Res. Pol. Pract. 39 (2), 221–236.

Hole, A.R., 2007. Estimating mixed logit models using maximum simulated likelihood. STATA J. 7 (3), 388–401.

Holmes, T.P., Boyle, K.J., 2005. Dynamic learning and context-dependence in sequential, attribute-based, stated-preference valuation questions. Land Econ. 81 (1), 114–126.

IPCC, 2014. In: Meyer, K.P.a.L.A. (Ed.), Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change R. IPCC, Geneva, Switzerland, pp. 151.

Johansson-Stenman, O., Svedsäter, H., 2008. Measuring hypothetical bias in choice experiments: the importance of cognitive consistency. B E J. Econ. Anal. Pol. 8.

Johansson-Stenman, O., Svedsäter, H., 2012. Self-image and valuation of moral goods: stated versus actual willingness to pay. J. Econ. Behav. Organ. 84 (3), 879–891.

Johnston, R.J., et al., 2017. Contemporary guidance for stated preference studies. J. Assoc. Environ. Resour. Econ. 4 (2), 319–405.

Kagel, J., Roth, A.E., 1995. The Handbook of Experimental Economics. PrincetonUniversityPress.

Kahneman, D., Knetsch, J.L., 1992. Valuing public goods: the purchase of moral satisfaction. J. Environ. Econ. Manag. 22, 57–70.

Kallbekken, S., Sælen, H., 2011. Public acceptance for environmental taxes: self-interest, environmental and distributional concerns. Energy Pol. 39 (5), 2966–2973.

Kragt, M.E., 2013. The effects of changing cost vectors on choices and scale heterogeneity. Environ. Resour. Econ. 54 (2), 201–221.

Ladenburg, J., Olsen, S.B., 2008. Gender-specific starting point bias in choice experiments: evidence from an empirical study. J. Environ. Econ. Manag. 56 (3), 275–285.

Lancaster, K.J., 1966. A new approach to consumer theory. J. Polit. Econ. 74 (2), 132–157.

List, J., Gallet, C., 2001. What experimental protocol influence disparities between actual and hypothetical stated values? Environ. Resour. Econ. 20 (3), 241–254.

List, J.A., et al., 2004. Examining the role of social isolation on stated preferences. Am. Econ. Rev. 94 (3), 741–752.

Lusk, J.L., Norwood, F.B., 2009. Bridging the gap between laboratory experiments and naturally occurring markets: an inferred valuation method. J. Environ. Econ. Manag. 58 (2), 236–250.

Lusk, J.L., Schroeder, T.C., 2004. Are choice experiments incentive compatible? A test with quality differentiated beef steaks. Am. J. Agric. Econ. 86 (2), 467–482.

Löschel, A., et al., 2013. The demand for climate protection—empirical evidence from Germany. Econ. Lett. 118 (3), 415–418.

McFadden, D., 1973. Conditional Logit Analysis of Qualitative Choice Behaviour. Academic Press, New York.

Murphy, J.J., et al., 2005. A meta-analysis of hypothetical bias in stated preference valuation. Environ. Resour. Econ. 30 (3), 313–325.

Mørkbak, M.R., et al., 2010. Choke price bias in choice experiments. Environ. Resour. Econ. 45 (4), 537–551.

Mørkbak, M.R., Olsen, S.B., 2015. A within-sample investigation of test–retest reliability in choice experiment surveys with real economic incentives. Aust. J. Agric. Resour. Econ. 59 (3), 375–392.

Nunes, P.A.L.D., et al., 2009. Decomposition of warm glow for multiple stakeholders: stated choice valuation of shellfishery policy. Land Econ. 85 (3), 485–499.

Nunes, P.A.L.D., Schokkaert, E., 2003. Identifying the warm glow effect in contingent valuation. J. Environ. Econ. Manag. 45 (2), 231–245.

Nyborg, K., 2000. Homo Economicus and Homo Politicus: interpretation and aggregation of environmental values. J. Econ. Behav. Organ. 42 (3), 305–322.

Poe, G.L., et al., 2005. Computational methods for measuring the difference of empirical distributions. Am. J. Agric. Econ. 87 (2), 353–365.

Poe, G.L., et al., 1994. Measuring the difference (X — Y) of simulated distributions: a convolutions approach. Am. J. Agric. Econ. 76 (4), 904–915.

Silbye, F., Sørensen, P.B., 2017. Subsidies to Renewable Energy and the European Emissions Trading System: Is There Really a Waterbed Effect? Danish Council on Climate Change.

StataCorp, 2013. Stata Statistical Software: Release 13. StataCorp LP, College Station, TX.

Svenningsen, L.S., 2017. Distributive Outcomes Matter: Measuring Social Preferences for Climate Policy. University of Copenhagen, Department of Food and Resource Economics.

Svenningsen, L.S., Thorsen, B.J., 2017. Preferences for Distributional Impacts of Climate Policy. University of Copenhagen, Department of Food and Resource Economics.

Swait, J., Louviere, J., 1993. The role of the scale parameter in the estimation and comparison of multinomial logit models. J. Market. Res. 30 (3), 305–314.

Train, K., 2009. Discrete Choice Methods with Simulation, second ed. Cambridge University Press.

Wiser, R.H., 2007. Using contingent valuation to explore willingness to pay for renewable energy: a comparison of collective and voluntary payment vehicles. Ecol. Econ. 62 (3–4), 419–432.