Can cheap talk scripts in combination with opt-out reminders nail down fat yes-tails in choice experiments?

Jürgen Meyerhoff* / Christine Bertram / Klaus Glenk / Katrin Rehdanz

* TU Berlin – juergen.meyerhoff@tu-berlin.de

Abstract

The problem of fat yes-tail responses is well known from contingent valuation but has not been investigated thoroughly in the context of choice experiments. In this study, we use eight independent split-samples with nearly 3600 respondents and systematically combine four different bid vectors with a joint cheap talk (CT) and opt-out reminder (OOR) device. Bid vectors differ with the respect to the four highest bid values increasing to usually not used values in choice experiments. Results clearly show that without a CT&OOR device a strong fat-tail effect is present, and WTP estimates are seriously inflated. In contrast, when the same bid vectors are used in combination with the CT&OOR device, the fat tail problem is strongly mitigated and WTP estimates are close to each other. However, results also indicate that the CT&OOR device does not nail down fat tails completely. As we have only applied CT and OOR jointly, more research into the effect of these devices and their strength is thus needed. Another striking result is that the impact of the CT&OOR device varies across attributes. We speculate that this is an effect of respondents' distance to the Baltic Sea and the varying non-use components of the attributes.

Keywords: choice experiment, fat yes-tails, cheap talk, opt-out reminder

1. Introduction

The phenomenon that some respondents to a stated preference survey accept even very high bid levels is known from the contingent valuation literature and called the fat tail problem. A consequence is that willingness to pay estimates are likely to be inflated and welfare measures biased. In this paper, we investigate whether the fat tail problem also exists when choice experiments (CE) are employed and, if so, whether a combination of cheap talk (CT) and an opt-out reminder (OOR) would “nail down” those fat tails. Fat tail problems have, to our knowledge, so far not systematically been investigated with respect to the application of choice experiments.

The motivation for this investigation are findings from a recent CE on the environmental quality of the Baltic Sea, which was meant to provide estimates for policy advice and where we found unreasonable high WTP estimates[1]. In order to come close to the choke price, we had used what we were thinking were clearly high levels for the highest bid: 800 Euros per year over a ten-year period. However, what happened was that the share of respondents accepting the high bid levels was much larger than expected, i.e., for the 800 Euro bid it was 19%. Accordingly, the marginal WTP estimates were large, and dropping choices with the highest bid resulted in significantly lower WTP estimates. An explanation for this is what Parson and Myers (2016), and others before them, call the fat yes-tail problem. They define a fat tail of a yes-response function in a contingent valuation study as a high and slowly declining yes-response rate at high bid levels.

If fat tail problems are similarly present in CE, the question arises what can be done to eliminate, or at least mitigate, this effect. Thus, another starting point for our analysis is Howard et al. (2017). They compare the effects of two devices to mitigate a hypothetical bias, a CT script and honesty priming, with a neutral control group in the context of choice experiments. CT scripts were introduced to the literature by Cummings and Taylor (1999) and are today one *ex ante* approach employed to reduce the hypothetical bias (Loomis 2014).[2] The main idea behind CT is to confront respondents explicitly with the problem of hypothetical bias by telling them that past surveys have shown that people tend to overestimate their actual willingness to pay.

Findings regarding the effectiveness of the CT script are mixed: in some cases, authors concluded that the CT script eliminated hypothetical bias, while others found that it reduces hypothetical bias or had even no effect (Johnston et al. 2016; Loomis et al. 2014). Howard et al. (2017) found in both an online choice experiment and in a CE conducted face-to-face greater sensitivity among respondents during choices made immediately after they had faced the CT script. However, they also observed that the more distant the choice is from having faced the CT script, i.e., the more choice

[1] Results are not yet published.

[2] Other *ex ante* approaches are to emphasize consequentiality, urge respondents to be honest, and reduce social desirability bias (Loomis 2014).

sets are between the text of the CT script and the choice set the respondents is answering, the more the effect of the CT script fades. One of their suggestions for future investigations is, thus, to identify means for extending the effectiveness of the CT script, among others, when respondents face multiple choice sets.

The problem that the effect of a CT script might fade when respondents go through a sequence of choice sets was also recognized by Ladenburg and Olsen (2014). They developed a so called opt-out reminder (OOR) that was presented to respondents on each choice set and used in combination with a CT script. The OOR points out on each set that if respondents think that the costs presented in the hypothetical alternatives are higher than the costs they would actually be willing to pay, then they should choose the status quo option, i.e., the alternative with no improvements compared to the current situation and thus a zero-price.

In view of the above, we test the following two hypotheses using data from a CE specifically designed for this study:

H1: Willingness to pay estimates are the same across independent samples with varying ranges of bid levels.

H2: Responses to the same bid vectors do not differ independently of whether respondents faced a CT script in combination with an OOR or not.

To test both hypotheses, we use eight independent samples with respondents facing the same questionnaire except for two issues, bid vectors and whether a device to mitigate the hypothetical bias was shown or not. In the first four samples, the bid vectors, including eight potential bid levels, vary. While the first four levels are identical across treatments, the last four levels vary, increasing to up to 1800 Euros in the fourth bid vector, an amount rarely used as annual payment in choice experiments. The remaining four samples mirror the bid vectors of the first four samples but respondents additionally face a combination of CT script and OOR before and while proceeding through the sequence of eight choice tasks. For comparison with our earlier study, we again focus on the environmental quality of the Baltic Sea in this CE.

We find that without CT script and OOR the fat yes-tail problem exists and marginal WTP estimates for some of our attributes are strongly inflated. However, when respondents faced the CT script in combination with the OOR, the willingness to pay estimates are significantly lower and much closer to each other across samples independent of the bid vector in a split sample. The CT&OOR device used has probably not completely nailed down fat tails, but seemingly reduced the fat tail problem significantly. The bid acceptance for the highest bid level fell by around 20%, and the marginal WTP estimates are close to each other across samples with CT&OOR while they clearly depend on the bid vector levels in the samples without CT&OOR.

2. Study design and survey method

2.1 Treatments and device to mitigate a hypothetical bias

Overall, we use eight independent split samples (Treatments T1 to T8). Figure 1 gives an overview about the treatments and Table 1 details the four different bid vectors. Half of the treatments received no CT&OOR device to mitigate the potential effects of the hypothetical nature of the survey (T1 to T4), while the remaining four treatments (T5 to T8) included a CT&OOR device.

Questionnaires were identical across the treatments except that bid vectors differ and that a subgroup of respondents received the CT&OOR device in order to remind them of the hypothetical nature of the survey. All respondents were informed at the beginning of the survey that results are intended to help decision makers to decide whether measures to increase the environmental quality of the Baltic Sea will be implemented (intended to promote consequentiality) and all respondents were, directly before they faced the choice sets, additionally reminded of potential consequences spending money for the Baltic Sea, i.e., that they would have to forgo other expenses. After this reminder those without CT&OOR treatment went to the sequence of choice tasks while the other group of respondents first faced the CT script before moving to the choice sets. The latter respondents also had the OOR on each choice task.

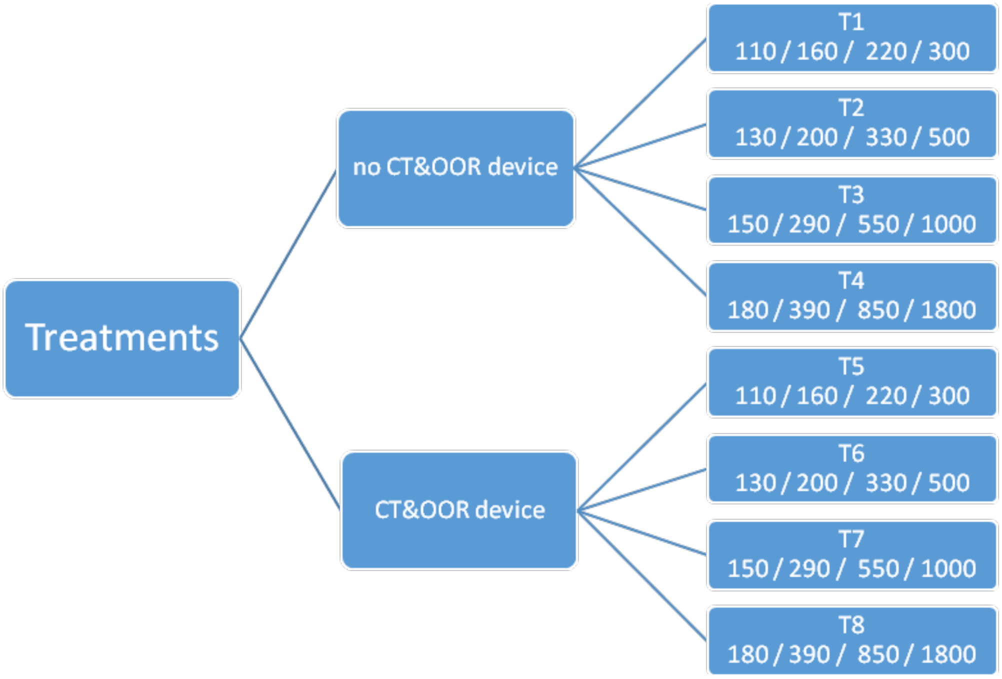


Figure 1: Overview over the treatments (levels of the four highest bids in Euro per year)

The payment respondents would have to make was introduced as a special tax for the Baltic Sea that all inhabitants would have to pay. Moreover, this tax would have to be paid for 10 years. Table 1 details the four bid vectors used. In all vectors, the first four bid levels have the same values, i.e., 10€, 25€, 50€, and 80€. This was motivated by the interest to see whether bid acceptance would also differ among the bids in the lower half of the bid vector. To test for potential fat tail effects, we increased the values of the bid levels for the highest four levels.

Compared to the levels Parson and Myers (2016) use, our highest value is rather modest: €1800 compared to US$ 10000. However, they asked for a one-time tax payment while in the current survey respondents were asked to pay the indicated amount annually for a period of 10 years, so undiscounted this would result in €18,000. As payment vehicle, a special tax for improving the environmental quality of the Baltic Sea was used.

Table 1: Bid vectors across treatments in Euro

Sample	Bid 1	Bid 2	Bid 3	Bid 4	Bid 5	Bid 6	Bid 7	Bid 8
T1 & T5	10	25	50	80	110	160	220	300
T2 & T6	10	25	50	80	130	200	330	500
T3 & T7	10	25	50	80	150	290	550	1000
T4 & T8	10	25	50	80	180	390	850	1800

Note: The first four levels are the same across all bid vectors.

To implement the combination of the CT script and the OOR, the CT&OOR device, in treatments T5 to T8 we used a light CT script. The original script used by Cummings and Taylor (1999) presented respondents a couple of paragraphs explaining what a hypothetical bias is, how this might affect their own responses and that people should actually think about their responses as if they would pay the stated amount of money in a real situation. However, such a lengthy explanation might not always be practical in surveys, especially in on-line surveys, and thus shorter scripts were tested (see Ladenburg and Olson 2014). We employ in this study a light version of a CT script that is much shorter than the original version. Moreover, our script is not neutral, i.e., it only reminds people of potentially overestimating their willingness to pay instead of neutrally saying that people might under- or overestimate their willingness to pay. The reason for this is that our concern, given the results of our previous policy oriented survey, had resulted in what we interpreted as unreasonable high WTP estimates. Following, to some extent, Ladenburg and Olson (2014), the wording of the CT script and the OOR in our survey is as follows:

> **<u>Cheap talk</u> script**
>
> In similar surveys, it was found that people tend to overestimate how much they are really willing to pay. When choosing among the following options, please bear in mind that an additional annual payment to protect the Baltic Sea will reduce your available

income. Depending on the amount of the annual payment for a program, you would have less money to spend on other expenses.

Opt-out reminder on each choice task

If the price for programs A and B is above the amount you would actually pay, then please choose "condition without further measures”

2.3 Econometric approach

As we are first of all interested in differences across the treatments, we use basic conditional logit (CL) models without investigating unobserved preference heterogeneity (see Howard et al. 2017). We compare CL models estimated for each treatment separately, and also estimate a CL model using a pooled data set. In this model, we try to capture differences across treatments by interacting all attributes with the four bid vectors and also incorporate an interaction term indicating whether the CT&OOR device was shown or not.

Generally, random utility theory assumes that the modeller does not possess complete information concerning the individual decision maker (subscript *n*). Thus, individual preferences are the sum of a systematic (*V*) and a random (*ε*) component

$$\mathrm{U_{ni} = V_{ni}(x_{ni}\beta) + \varepsilon_{ni}} \quad (1)$$

where *U_{ni}* is the true but unobservable utility associated with alternative *i* out of a set of available alternatives, V_{ni} is the measurable or deterministic part which itself is a function of the attributes (x_{ni}), β is a vector of coefficients reflecting the desirability of the attributes, and ε_{ni} is a random term with a zero mean. This error term represents attributes and characteristics unknown to the researcher, measurement error and/or taste heterogeneity among respondents. Selection of one alternative over another implies that the utility (U_{ni}) of that alternative is greater than the utility of the other alternatives:

$$\mathrm{P(i) = Prob(V_i + \varepsilon_i > V_j + \varepsilon_j) \quad \forall\, j \in C, j \neq i} \quad (2)$$

Assuming that the error components are distributed independently and identically (IID) following a type 1 extreme value distribution, one gets the multinomial logit (MNL) model where the probability of individual *n* choosing alternative *i* takes the form:

$$P_{ni} = \frac{\exp(\mu V_{ni})}{\sum_{j \in C} \exp(\mu' V_{nj})} \quad (3)$$

where μ is a scale parameter which is commonly normalised to 1 in practical applications for any one data set as it cannot be identified separately from the vector of parameters. The scale parameter is inversely proportional to the error variance σ_ε^2:

$$\mu = \frac{\pi}{\sqrt{6\sigma_\varepsilon^2}} \quad (4)$$

The assumption of a constant error variance across individuals has been questioned and a heteroskedastic logit model was suggested as an alternative (HL; e.g., Swait and Louviere, 1993). Here the scale parameter is no longer a constant term as it allows for unequal variances across unobserved components from two or more data sources. Whether the variation in scale across data sources can be explained by factors such as respondent characteristics or the design dimensions of the choice sets is investigated using the following HL expression (Caussade et al. 2005; DeShazo and Fermo 2002):

$$P_{ni} = \frac{\exp(\mu_n V_{ni})}{\sum_{j \in C} \exp(\mu_n V_{nj})} \tag{5}$$

where μ_n = $\exp(\gamma' Z_n)$ with $\boldsymbol{Z_n}$ a vector of respondent specific characteristics including the design dimensions a respondent is randomly assigned to and γ' a vector of parameters indicating the influence of those characteristics on the error variance (Hole 2006). The exponential form ensures a positive scale factor. In the heteroskedastic logit model used here the scale parameter is specified as a function of the differences among cost bids and of the maximum willingness to pay (Dellaert et al. 1999). The parameters γ' and β' are jointly estimated via maximum likelihood using the Stata program *clogithet* (Hole 2006).

3. Survey design

3.1 Attributes and experimental design

The attributes and their levels (Table 2) were developed through expert interviews in the context of a project regarding the environmental quality of the Baltic Sea as well as by resorting to previous valuation studies concerned with marine environmental quality (e.g., Norton and Hynes, 2014). They aimed to describe different aspects of achieving a good environmental status of the Baltic Sea, in accordance with the EU Marine Strategy Framework Directive (MSFD). The first attribute, water clarity, describes how far one can see below the water surface. This is an attribute often used to describe overall water quality in environmental valuation studies. The second attribute refers to the state of the fish stocks in the sea, i.e., whether fish stocks are overfished or stable. This attribute represents a major direct pressure on marine living resources induced by human use. The third attribute, biodiversity, was chosen to describe the overall state of the marine ecosystem, including all living species. The fourth attribute refers to the impact of coastal protection measures on the landscape, which may be an important factor for the perceived quality of a visitor's stay at the sea, and the fifth attribute is the amount of litter that is present on the beaches and in the water column. The sixth and last attribute is the cost attribute, described as the yearly expenses per person for a "Baltic Sea tax".

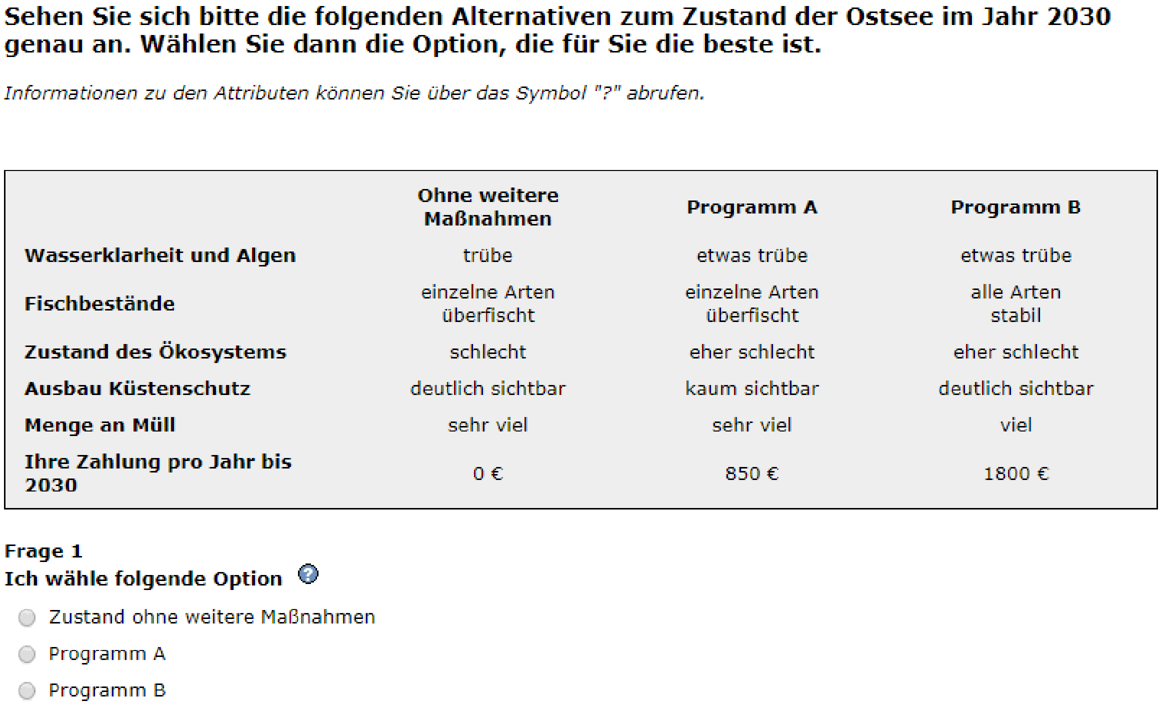
Table 2: Attributes and levels

Attribute	Levels

Water clarity	*Turbid*, rather turbid, rather clear, clear
Fish stocks	*Some species overfished*, stable fish stocks
State of biodiversity	*Bad*, rather bad, rather good, good
Coastal protection	*Mainly strong impact*, mainly low impact (on landscape)
Amount of litter	*Very much*, much, little, very little
Cost	*Depending on cost vector (see Table 1)*

Note: Attribute levels of the status quo (SQ) alternative are in italics.

We used a Bayesian efficient design with uniform priors for the attribute parameters to allocate the attribute levels to the unlabelled alternatives. The range for each prior was determined based on pilot surveys and also using results from the main survey in the same project. The D-efficiency design criterion for a multinomial logit model was used. To allow for uncertainty in the value of the prior, modified Latin hypercube sampling (ChoiceMetrics, 2012) was applied, and 1,000 draws were taken for each parameter prior to uniform distributions. The algorithm stopped when no improvements with respect to the optimization criteria were found within 60 minutes. The final design comprised 32 tasks for each treatment distributed to four blocks with eight choice tasks. We used the same experimental design for all 8 treatments. This might have resulted in a loss of efficiency but given the target sample sizes of 400 respondents per split we were not strongly concerned. Figure 2 shows an example choice set excluding the CT&OOR device, the zero price (SQ) alternative is presented first followed by the two hypothetical alternatives.



Sehen Sie sich bitte die folgenden Alternativen zum Zustand der Ostsee im Jahr 2030 genau an. Wählen Sie dann die Option, die für Sie die beste ist.

Informationen zu den Attributen können Sie über das Symbol "?" abrufen.

	Ohne weitere Maßnahmen	Programm A	Programm B
Wasserklarheit und Algen	trübe	etwas trübe	etwas trübe
Fischbestände	einzelne Arten überfischt	einzelne Arten überfischt	alle Arten stabil
Zustand des Ökosystems	schlecht	eher schlecht	eher schlecht
Ausbau Küstenschutz	deutlich sichtbar	kaum sichtbar	deutlich sichtbar
Menge an Müll	sehr viel	sehr viel	viel
Ihre Zahlung pro Jahr bis 2030	0 €	850 €	1800 €

Frage 1
Ich wähle folgende Option

○ Zustand ohne weitere Maßnahmen
○ Programm A
○ Programm B

Figure 2: Example choice task without OOR

3.2 Survey and sample

The questionnaire started with questions concerning the respondents' familiarity with the Baltic Sea, e.g., whether they had visited the Baltic Sea in the past and, if so, how frequently. The individuals were then introduced to the problems the Baltic Sea faces today and will very likely face in the future, and they were informed that suitable management actions could influence the future state of the Baltic Sea. Afterwards, they were introduced to the attributes. Respondents were then randomly assigned to one of the eight treatments, and the choice tasks were presented to them in a random order. After the section with the choice sets, respondents faced a couple of debriefing questions, among them a ranking of the attribute importance for their decisions a battery of items concerning respondents' decision style.

In total, we aimed at least at 400 completed interviews per treatment. Respondents were drawn from a nationwide online-panel in Germany and were invited to participate in the survey. Interviews were conducted in June/July 2017. In the end, 3576 useable interviews were gained. The number of respondents varies per treatment from 414 (T1) with the lowest number of participants to 479 (T2) with the highest number of respondents. Table 3 reports descriptive statistics for all 8 split samples as well as the overall mean. The means per split for the variables gender, age, whether people had visited the Baltic Sea within the last twelve months or are certain to go there within the next 12 months, people per household and a rating of their financial situation (on a four-point scale from very good to very bad) indicate that there are no systematic differences between the 8 split samples. The mean for each split is generally close to the overall mean.

Table 3: Descriptive statistics of sample

Split	N	Gender (female=1)	Age (years)	Visit last 12 months (mean)	Intended visit next 12 months (mean)	People per household (mean)	Financial situation* (mean)
T1	414	0.54	49.34	0.21	0.31	2.21	2.47
T2	479	0.53	50.06	0.20	0.30	2.21	2.50
T3	444	0.50	50.82	0.19	0.34	2.15	2.49
T4	439	0.53	49.34	0.20	0.32	2.17	2.51
T5	477	0.55	50.20	0.20	0.30	2.17	2.50
T6	431	0.54	48.22	0.22	0.31	2.26	2.59
T7	449	0.50	50.55	0.20	0.30	2.19	2.57
T8	443	0.54	49.24	0.21	0.32	2.33	2.50
Total	3576	0.53	49.74	0.20	0.31	2.21	2.52

Note: * measured on a four-point scale

4. Results

The presentation of the results starts with a comparison of the share of status quo choices across treatments. This is followed by an analysis of the bid acceptance curves for each treatment comparing them for the same bid vector levels but with or without CT&OOR device. Next, we move to the presentation of results from the conditional and heteroscedastic logit models, before the final sections is about the WTP estimates across treatments. Finally, we present the responses to the choice certainty questions and WTP estimates from a CL model using only respondents who had responded a minimum level of certainty.

4.1 Status quo choices

As a first step in our analysis, we investigate whether, and if so, to what extent the increasing bid levels and the CT& OOR device impact on respondents' SQ-choices. Table 4 shows for all splits the share of respondents who have chosen the SQ-option and how many times they did so. Starting with the four splits without the CT&OOR device, we find that clearly more than 50% of all respondents have never chosen the SQ-option (column "0" in table 4), i.e., those respondents indicated on all choice sets that they are willing to pay a positive amount. In contrast, only 4 to 6 percent of the respondents have never chosen an alternative with a positive price (column "8" in table 4) indicating that they definitively do not prefer to improve the environmental conditions of the Baltic Sea when they would have to give up money for this.

Looking at the four splits where respondents faced the CT&OOR device, there is a clear change in the share of respondents who choose the SQ-option as well as the number of times they did so. In all four splits at most 25 percent have never chosen the SQ option, this is more than half of the share observed in the splits without device. At the opposite side of the range, we see that the share of those who never choose an alternative with a positive price has doubled.

Table 4: Share of respondents choosing the status quo option a certain time (in %)

Treatment		*Number of times the status quo options was chosen*									*Total*
		0	1	2	3	4	5	6	7	8	
No device	1	64	11	13	4	2	1	1	0	5	100
	2	59	14	13	4	2	1	1	1	4	100
	3	61	17	10	2	3	2	1	1	4	100
	4	54	18	14	3	2	2	1	1	6	100
CT& OOR	5	25	15	18	12	7	5	5	2	9	100
	6	23	14	23	7	7	6	6	4	10	100
	7	18	16	22	11	5	5	6	3	13	100
	8	20	15	23	12	7	5	5	3	9	100

Note: Treatment 5 to 8 presented respondents cheap talk scripts and opt-out reminders

4.2 Bid acceptance by treatment

Figure 2 shows bid acceptance as a yes-response function in percent by bid vector comparing whether respondents faced the CT&OOR device or not. It is striking that the functions for the samples without the CT&OOR device do not show a clearly downward slope. Especially for the first bid vector with a maximum of 300 €, the percentage of acceptance does not go below 35% for any bid, not even for the 300€ bid. Looking at the other bid vectors, the acceptance for the highest bid is around 10 percentage points lower as for the lowest bid, not a really strong downward slope. Interesting to note is also that the acceptance for the first bid in all four samples without CT&OOR device is around 40%, a remarkably low value given that the first bid is in all samples only €10. Another clear pattern is that all four functions for the yes-response without CT&OOR device have a clear peak at the €80 bid, the highest bid value that all vectors have in common. Acceptance for this value is at the same time also the highest for all bid values. Moving to the higher bid values, we find another pattern across all these four samples: the first bid value that varies across samples, lying between €110 (T1) and €180 (T4) have clearly lower acceptance values, while the following levels show in all four samples clearly higher acceptance rates although the values being obviously higher (T1: €160; T4: €390). It looks as if respondents associate a kind of suitability with certain bid levels as these levels express the “right” willingness to pay for improving the quality of the Baltic Sea.

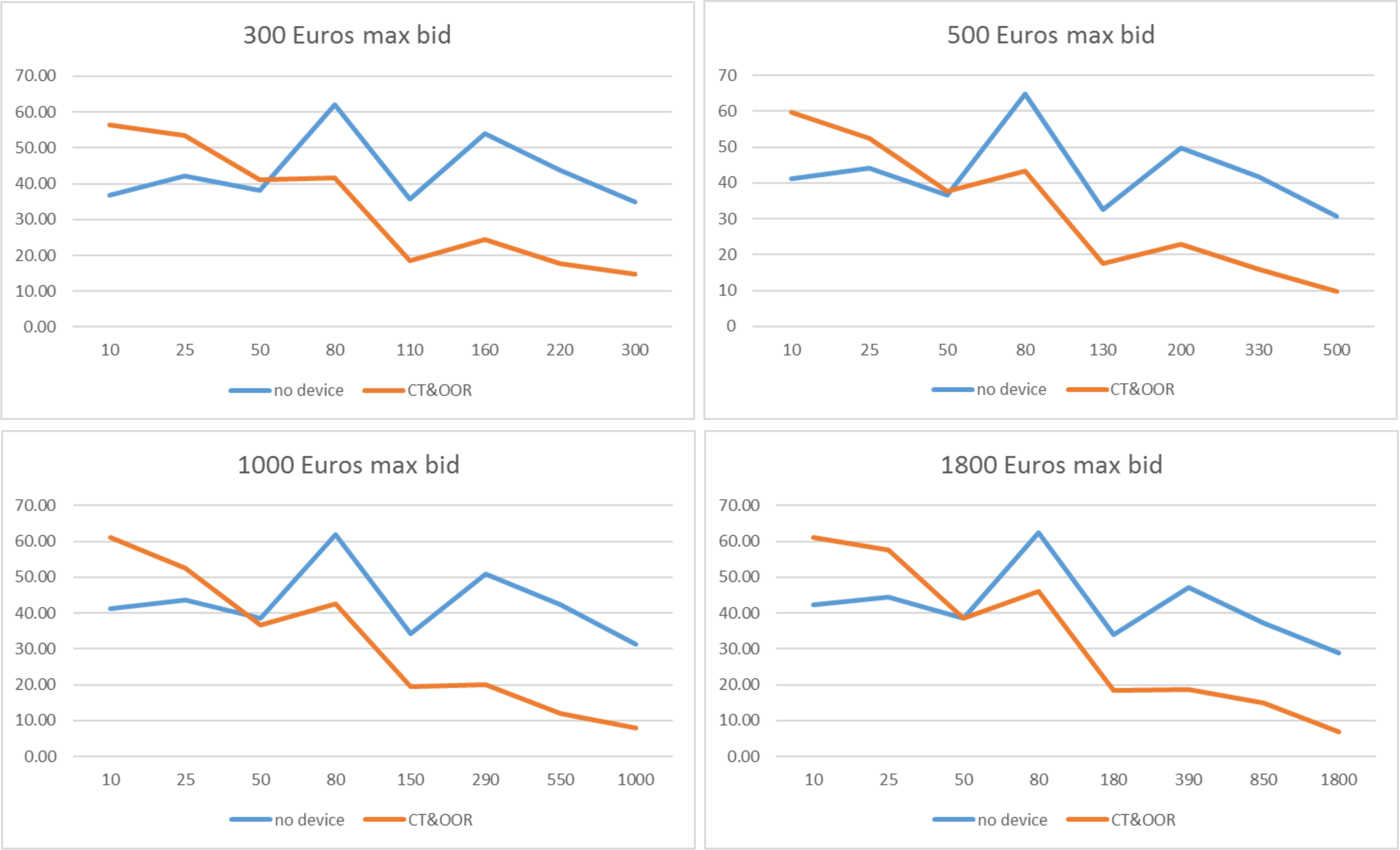


Figure 2: Bid acceptance in percent by bid vectors and treatments

This picture clearly changes when we look at the yes-response functions from the samples with the CT&OOR device. Here, in all four cases, the acceptance for the lowest bid (10€) is much higher, around 60% and thus 20percentage points higher than in the samples without the CT&OOR device. The yes-response function from these four samples also all show a clearer downward slope. However, we also find peaks in the function but compared to the samples without the CT&OOR device they are not that clear and do not occur in all four samples at the same bid values. For example, apart from sample T1 the remaining samples have a peak at €80, and T5 and T6 have a small peak at the sixth bid level. Another striking observation is that acceptance for the €50 bid is very similar across all eight samples and lies always close to 40%.

4.3 Choice models

Table 5 gives the estimates of the conditional logit models separately for each treatment. Focussing of similarities and dissimilarities, we firstly see that fish stocks, biodiversity, cost and the ASCsq are all highly significant across treatments and have the same sign. The probability that an alternative is chosen increases when fish stocks are stable and biodiversity is in a better condition while higher levels of the cost attribute have a negative impact on choosing an alternative. The ASCsq indicates that, on average, people want to move away from the current situation and prefer high quality levels of the Baltic Sea. In contrast, the attributes litter and water clarity are not in all treatments as relevant for peoples' choices as the previously mentioned attributes. Litter has a t-value of 0.26 (i.e., a high standard error) in T8, the treatment with the highest two bid levels. Water clarity has low t-values in the treatments T6 to T8. An explanation might be that respondents value improvements in the attributes presented on the choice sets differently when they are explicitly confronted with the problem of a potential hypothetical bias (face the cheap talk script) and when bid levels increase. We will come back to this point in the concluding section of the paper.

Table 5: Conditional logit models per treatment (|t-values| in parenthesis)

	T1	T2	T3	T4	T5	T6	T7	T8
Water clarity	0.220	0.129	0.133	0.059	0.150	0.003	-0.008	-0.034
	(9.99)	(6.55)	(6.89)	(3.07)	(6.78)	(0.13)	(0.33)	(1.49)
Fish stocks	0.372	0.365	0.385	0.417	0.271	0.186	0.187	0.182
	(9.20)	(9.69)	(9.92)	(10.62)	(6.33)	(4.04)	(4.12)	(4.07)
Biodiversity	0.366	0.295	0.258	0.232	0.221	0.122	0.053	0.046
	(16.98)	(15.56)	(13.77)	(12.57)	(10.23)	(5.39)	(2.41)	(2.16)
Coastal	0.030	-0.030	-0.011	-0.004	-0.073	-0.049	-0.075	-0.161
	(0.73)	(0.80)	(0.28)	(0.10)	(1.68)	(1.05)	(1.60)	(3.52)
Litter	0.391	0.354	0.314	0.239	0.182	0.112	0.046	0.006
	(17.77)	(18.65)	(16.48)	(12.96)	(8.32)	(5.00)	(2.11)	(0.26)
Cost	-0.006	-0.004	-0.002	-0.001	-0.010	-0.006	-0.003	-0.002
	(15.35)	(17.83)	(16.91)	(17.22)	(25.34)	(23.85)	(24.39)	(23.79)
ASCsq	-0.400	-0.490	-0.551	-0.457	-0.161	-0.330	-0.378	-0.583
	(4.86)	(6.51)	(7.02)	(5.88)	(2.07)	(3.89)	(4.45)	(6.95)
ASC2	0.053	0.089	0.041	0.096	0.108	0.164	0.106	0.099
	(1.37)	(2.45)	(1.09)	(2.52)	(2.61)	(3.64)	(2.36)	(2.24)
Observations	9936	11496	10656	10536	11448	10344	10776	10632

Absolute *t* statistics in parentheses
$p < 0.05$, $p < 0.01$, $p < 0.001$

Table 6 presents the results from the models estimated using the pooled data set. While model 1 is a basic CL model, models 2 and 3 incorporate separately the differences across the treatments, i.e., the different bid vectors or the presence of the CT&OOR device, by interaction effects. Model 4 has both differences captured in one model and, finally, model 5 (Heteroskedastic logit) additionally accounts for scale differences caused by the different bid values across bid vectors and the maximum value of the bid value in a bid vector (see Dellaert et al. 1999).

Starting with the basic CL model, all parameters are as expected except that *Clarity* has only low relevance and that *Coastal protection* has a negative sign. Both unexpected results are, as we will discuss below, are rather due to the pooled data set and not accounting for differences across treatments in the basic CL model (compare also to the results presented in Table 5). In line with expectations are the findings that respondents are in favour of stable *Fish stocks*, higher levels of *Biodiversity* and less *Litter* on the beaches and in the water column. In contrast, cost has a negative sign making costlier alternatives less likely to be chosen. The negative sign of the ASC for the status quo suggests that people prefer, on average, to increase the quality of the Baltic Sea, and the ASC for the second alternative is positive. Respondents have a positive tendency to choose Programme A over Programme B. Both ASCs remain significant and have the same sign across all models.

Table 6: Conditional logit models estimated for the pooled data set

	(1)	(2)	(3)	(4)	(5)
	Basic CL	Model 1 plus BidV	Model 1 plus CT	Complete CL	Heteroskedastic logit
ASCsq	-0.356	-0.268	-0.494	-0.429	-0.351
	(13.30)	(10.03)	(17.42)	(15.17)	(9.69)
ASC2	0.082	0.085	0.089	0.091	0.100
	(5.87)	(6.10)	(6.26)	(6.36)	(6.26)
Clarity	0.006	0.163	0.056	0.239	0.256
	(0.88)	(11.97)	(6.24)	(15.61)	(15.24)
Clarity * bidvector 2		-0.091		-0.099	-0.095
		(4.95)		(5.27)	(4.74)
Clarity * bidvector 3		-0.089		-0.103	-0.088
		(4.95)		(5.57)	(4.36)
Clarity * bidvector 4		-0.127		-0.133	-0.098
		(7.09)		(7.20)	(4.74)
Clarity * CT & OOR			-0.100	-0.148	-0.131
			(7.98)	(11.71)	(9.21)
Fish stocks	0.219	0.290	0.310	0.404	0.436
	(15.55)	(10.60)	(16.56)	(13.06)	(12.98)
Fish stocks * bidvector 2		-0.004		-0.020	-0.016
		(0.12)		(0.51)	(0.39)
Fish stocks * bidvector 3		-0.010		-0.009	-0.005
		(0.27)		(0.22)	(0.11)
Fish stocks * bidvector 4		0.047		0.038	0.047
		(1.21)		(0.97)	(1.07)
Fish stocks * CT & OOR			-0.183	-0.219	-0.224
			(6.63)	(7.87)	(7.28)
Biodiversity	0.129	0.267	0.217	0.381	0.404
	(19.34)	(20.09)	(25.02)	(25.41)	(24.37)
Biodiversity * bidvector 2		-0.054		-0.067	-0.061
		(3.04)		(3.70)	(3.12)
Biodiversity * bidvector 3		-0.105		-0.118	-0.102
		(6.04)		(6.57)	(5.23)
Biodiversity * bidvector 4		-0.107		-0.114	-0.075
		(6.16)		(6.35)	(3.71)
Biodiversity * CT & OOR			-0.175	-0.215	-0.199
			(14.42)	(17.71)	(14.59)

Coastal protection	-0.122	-0.043	-0.079	0.015	0.029
	(8.59)	(1.55)	(4.18)	(0.47)	(0.87)
Coastal protection * bidvector 2		0.006		0.002	0.007
		(0.16)		(0.05)	(0.16)
Coastal protection * bidvector 3		0.001		0.001	0.004
		(0.04)		(0.02)	(0.10)
Coastal protection * bidvector 4		0.000		0.000	0.009
		(0.01)		(0.00)	(0.21)
Coastal protection * CT&OOR			-0.092	-0.127	-0.126
			(3.31)	(4.54)	(4.06)
Litter	0.132	0.258	0.254	0.405	0.430
	(19.78)	(19.12)	(29.13)	(26.62)	(25.46)
Litter * bidvector 2		-0.015		-0.034	-0.026
		(0.82)		(1.87)	(1.35)
Litter * bidvector 3		-0.078		-0.086	-0.066
		(4.43)		(4.78)	(3.34)
Litter * bidvector 4		-0.116		-0.124	-0.083
		(6.62)		(6.90)	(4.03)
Litter * CT&OOR			-0.240	-0.278	-0.266
			(19.90)	(22.98)	(19.57)
Cost	-0.002	-0.007	-0.001	-0.007	-0.008
	(44.54)	(28.74)	(26.75)	(27.15)	(24.34)
Cost * bidvector 2		0.003		0.003	0.003
		(10.15)		(9.97)	(8.42)
Cost * bidvector 3		0.005		0.005	0.005
		(19.03)		(19.11)	(16.17)
Cost * bidvector 4		0.006		0.006	0.006
		(23.24)		(23.06)	(18.71)
Cost * CT&OOR			-0.002	-0.001	-0.002
			(18.70)	(12.35)	(10.51)
Heteroskedasticity					
Differences cost bids					0.0001
					(2.77)
Maximum cost bid					-0.0001
					(8.82)
Observations	85824				
Respondents	3576				

Model 2 adds to the basic CL the interactions with the bid vectors measuring differences to the first bid vector. For the non-monetary attributes *Clarity*, *Biodiversity*, and *Litter*, the interactions are each time significant and have a negative sign, i.e., in case of bid vectors with higher bid values utilities from an improvement decrease. Moreover, we find that the main effect for *Clarity* is significant after incorporating the differences among bid vectors. In case of the attribute *Fish stocks* the interactions with the bid vectors are not significant as indicated by the low t-values. Regarding this attribute, the respondent's utility does not seem to be affected by rising bid values. For coastal protection, neither the main nor the interaction effects are significant. In contrast, for the cost attribute interactions suggest decreasing cost sensitivity, i.e., respondents accept higher cost values in order to choose other than the zero-price alternative.

Model 3 includes the interaction between all attributes and whether respondents faced the CT&OOR device. The uniform effect is that for all attributes the parameters significantly decrease, i.e., the interaction term has a negative sign. Utilities from quality improvements decrease while cost sensitivity increases when the CT&OOR device was implemented. Interestingly, this applies even to the attribute *Coastal protection*. Moving to Model 4 incorporating both changing bid vectors and the presence of the CT&OOR device, we find that the effects are very stable and signs and significances remain.

The heteroskedastic logit model (Model 5) shows that both differences among bid values across treatments and the varying value of the highest bid of each bid vector significantly impact on scale. Larger differences among bids increase scale and thus lower the error variance while higher bid values have the opposite effect. The higher the value of the highest bid is the more the error variance increases.

4.4 Willingness to pay estimates

Figure 3 shows the mean willingness to pay estimates by treatment resulting from the heteroskedastic logit model (Model 5) in Table 6. The values are also reported in Table 7 showing also the standard errors for each WTP estimate. We only report here the values for four attributes as the attribute *coastal protection* varies strongly regarding its relevance across the different models and is thus less informative for comparisons. The estimates were calculated based on the corresponding combination of main effects and interactions as reported in Table 6. Varying across attributes, we find for the first four treatments a very clear pattern. When the values of the highest four bids in each bid vector increase, the mean WTP estimates increase as well. Thus, the fat tails strongly influence mean WTP estimates. The range between the lowest and the highest WTP values varies by attribute, and is especially obvious for the attributes *Fish stocks*, *Biodiversity*, and *Litter*. Interestingly, all three attributes have clear non-use value components compared to *Clarity* and *Coastal protection* that might be more relevant for users.

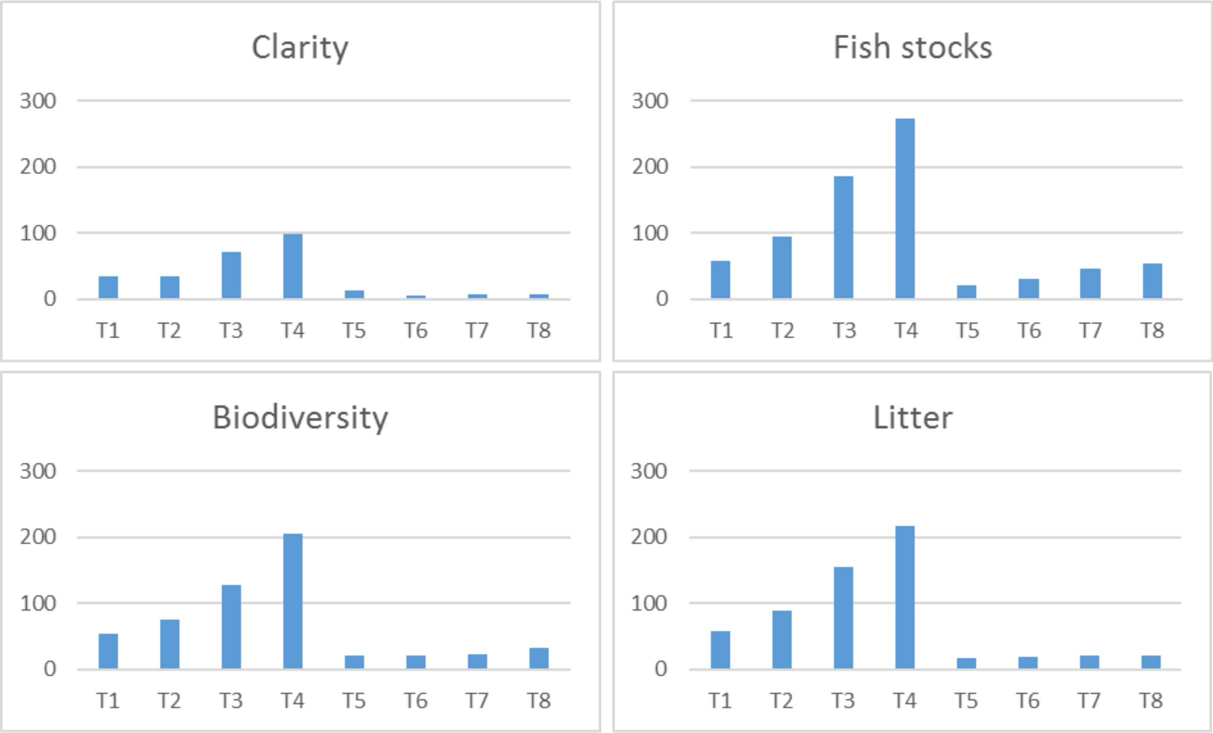


Figure 3: Bar graph of mean WTP estimates from conditional logit model

Findings look very different when relying on the four samples where respondents faced the CT&OOR device (T5-T8). Here, we find no similar pattern to the first four samples, i.e., mean WTP values do increase when the highest four values of the bid vectors increase but to a much lesser extent. For the attribute water clarity, we get very small WTP values, clearly below 10€ per year for the treatments T6 to T8. This might be another indication that the willingness to pay for water clarity is mainly driven by use-values, and facing the CT&OOR device improvements in this attribute seems to become less important. The mean estimates for the remaining attributes are on a similar level, following, as already described, the increases of the bid vector values to some extent but not comparable to the treatments without CT&OOR device. However, both the WTP estimates for *Fish stocks* and *Biodiversity* increase in T8, the treatment with the highest bid value of the *Cost* attribute.

Table 7: Marginal WTP estimates based on model 5 (Table 6)

	Clarity		**Fish stocks**		**Biodiversity**		**Litter**	
	mean	s.e.	mean	s.e.	mean	s.e.	mean	s.e.
T1	34.05	1.88	57.96	4.56	53.79	2.15	57.22	2.21
T2	35.15	3.07	95.25	8.16	75.15	3.44	88.27	3.72
T3	71.46	6.12	185.01	16.74	128.36	7.24	154.71	8.12
T4	98.79	9.53	273.34	27.31	206.31	13.17	217.98	13.48
T5	12.70	1.54	21.50	3.37	20.80	1.45	16.62	1.48
T6	4.35	2.46	30.65	4.97	20.89	2.17	19.88	2.15
T7	7.99	3.63	45.14	7.41	21.94	3.30	20.87	3.28
T8	6.79	4.68	53.86	9.06	32.93	3.99	20.62	4.14

4.5 Choice certainty

After respondents had finished the sequence of choice tasks they were asked to indicate their choice certainty. This question is discussed in the literature as an *ex-post* device to calibrate WTP estimates (see Champ et al. 1997). Table 8 presents the mean and the median for the responses to the 10-point certainty scale. Overall, the mean values are similar across splits. The overall mean is 6.72 and the median is always 7. Mean values are slightly lower for the samples where respondents faced the CT&OOR device. Using t-tests for all possible combinations reveals that the mean values of T7 and T8 are different from the first four treatments (T1 to T4) using conventional levels. Thus, the results suggest that stated choice certainty is not strongly affected by the different bid vectors, i.e., increasing bid levels, and is also only slightly affected by the CT&OOR device, i.e., certainty decreases marginally.

Table 8: Overall certainty about choices across treatments

Treatment	N	Mean	Sd (mean)	Median
T1	414	6.89	1.95	7
T2	479	6.81	1.90	7
T3	444	6.91	1.91	7
T4	439	6.83	1.94	7
T5	477	6.68	2.03	7
T6	431	6.67	2.03	7
T7	449	6.53	1.95	7
T8	443	6.48	2.12	7
Total	3576	6.72	1.98	7

Table 9 gives the WTP estimates from the same model using pooled data but only for those respondents who stated a value of seven or higher on the certainty scale, indicating a level of certainty that has been used in other studies as a value to calibrate WTP estimates (see Morrison & Brown, 2009). However, using the reduced sample does not lead to convergence between the treatments with and without CT&OOR device. The estimates from the treatments with CT&OOR are obviously on a different, much lower, level, and much closer to each other than from the treatments without CT&OOR.

Table 9: Marginal WTP estimates based on Model 5 but with choice certainty > 6

	Clarity		Fish stocks		Biodiversity		Litter	
	mean	s.e.	mean	s.e.	mean	s.e.	mean	s.e.
T1	40.18	2.79	69.27	6.77	69.27	3.53	73.35	3.82
T2	45.72	4.74	119.47	13.20	119.47	6.41	125.36	7.49
T3	97.00	9.44	229.20	26.34	159.33	11.98	207.59	14.81
T4	136.65	14.64	342.98	41.86	279.62	22.30	279.58	22.17
T5	14.21	2.27	22.82	4.98	25.82	2.13	21.66	2.14
T6	5.20	3.88	34.03	7.74	31.40	3.33	31.53	3.31
T7	12.56	5.64	50.39	11.60	26.58	5.18	29.46	5.12
T8	11.60	7.29	60.98	14.43	49.60	6.23	22.50	6.66

Note: Number of observations is 50856

5. Discussion and conclusions

With respect to the two hypotheses stated at the beginning of the paper we find two striking results: Firstly, WTP estimates are clearly different across the four samples when the highest bid vector levels increase and respondents do not face the CT&OOR device. Thus, as reported by Parson and Myres (2016) for CVM, we also find fat yes-tails in our CEs when the bid vector increases. The bid acceptance for the highest bid is in the four samples around 30% (T1 has even more with 35% for the highest value, 300 Euro), more than one would expect given the high values of up to 1800 Euro per year for a ten-year period.

Secondly, the WTP values decrease significantly in the split samples with the CT&OOR device. Overall, bid acceptance is 20percentage points lower in the mirroring sample with CT&OOR. From the split with the highest bid level we have a bid acceptance of 6.9% compared to 28.8% without the CT&OOR device. Thus, with respect to our second hypothesis we find that the combination of CT plus an OOR has a significant effect and we can reject the second hypothesis.

However, as the results show as well, the combination of the CT & the OOR reduces the fat yes-tail occurrence but does not eliminate it completely. WTP estimates are more similar to each other in the samples with the CT&OOR device, and lower than in treatment T1 with the lowest bid level values, but we still observe slightly increasing WTP estimates for some attributes when the higher values of the bid vectors increase. Moreover, we only employed a combination of CT and OOR and thus cannot disentangle whether the observed effects are due to one or both devices or whether indeed the combination of both was necessary to get the effects we found. Future studies might thus devote more resources to investigating whether the clear effect we observe indeed requires the combination of both tools or whether one of them could be sufficient as well. Given the evidence in the literature our best guess is that the combination of both instruments makes the difference, but we have not investigated this here.

Overall, the WTP estimates of the samples with the CT&OOR device seem to be reasonable, however, we cannot rule out that our device overcorrected, something that was already found in the literature. As we do not have real payments for the environmental changes in question we cannot assess whether we have eliminated the hypothetical bias. Recalling the statement by Harrison (2006) that devices such as cheap talk scripts are not a magic bullet this is rather unlikely. However, WTP estimates of the four splits with CT&OOR result in fairly equal estimates, clearly in contrast to the four splits without that device.

Results suggest that the effect of the CT&OOR device is different for the different attributes. Preferences for stable fish stocks and the state of biodiversity seem to be less effected than preferences for water clarity and litter on the beach and in the water column. A reason for this might be respondents geographical distance to the Baltic Sea that varies due to the national samples we used. Respondents living further away, and thus having a lower probability to visit the Baltic Sea, might be less concerned by bad water clarity levels and more litter than respondents that are living close. As we collected the postal codes for each respondent we at least can calculate a rough estimate for the distance between their place of residence and the Baltic Sea. Additionally, we have information about respondents past visits to the Baltic Sea and also test to what extent the frequency of past visits has an influence on the importance of certain attributes, e.g., water clarity that might become more important the more often people visited the sea.

A limitation could be that so far we have only accounted for observed heterogeneity due to the treatments and have not accounted for unobserved taste heterogeneity. Accordingly, models also capturing unobserved heterogeneity would show whether the differences between WTP estimates narrow when unobserved heterogeneity is recognized. Given the strong effect of the bid values we found in the treatments without CT&OOR device we do not expect that accounting for unobserved taste heterogeneity would eliminate this effect, but it might lower it. Regarding other future steps, we think about investigating the influence of the order respondents faced the different bid values. The reason for this is that yes-responses increase for higher bid values while they are comparable low for low bid values. A hypothesis is that respondents in the treatments without CR&OOR device, who have seen other values than the lowest bid first, may think that the lowest bid is too low as an expression of value for the Baltic Sea. If no CT&OOR is present, this might promote thinking that accepted bids do not represent money they would have to pay later but is rather a mean to express their "value" for the Baltic Sea, and low bid values do not reflect what people think is fair value for the good in question. Finally, another step would be to estimate mixed logit model with flexible distributions (Train 2017). These models might be suitable to gain more insights into the fat yes-tails and to what extent they are nailed down when the CT&OOR device is applied

References

- Champ, P. A., Bishop, R. C., Brown, T. C., & McCollum, D. W. (1997). Using Donation Mechanisms to Value Nonuse Benefits from Public Goods. *Journal of Environmental Economics and Management* 33, 151-162.
- ChoiceMetrics, 2014. Ngene 1.1.2. User manual & reference guide. Sydney, Australia.
- Cummings, Ronald G., Taylor, Laura O. (1999). Unbiased value estimates for environmental goods: A cheap talk design for the contingent valuation method. *American Economic Review* 89:649–65.
- Dellaert, Benedict G. C., Brazell, Jeff D., Louviere, Jordan J. (1999). The Effect of Attribute Variation on Consumer Choice Consistency. *Marketing Letters* 10 (2): 139-147.
- Harrison, G. W. (2006). Experimental Evidence on Alternative Environmental Valuation Methods. *Environmental and Resource Economics* 34(1), 125-162.
- Hole A R (2006) Small-sample properties of tests for heteroscedasticity in the conditional logit model. *Economics Bulletin* 3: 1-14.
- Howard, G., et al. (2017). Hypothetical Bias Mitigation Techniques in Choice Experiments: Do Cheap Talk and Honesty Priming Effects Fade with Repeated Choices? *Journal of the Association of Environmental and Resource Economists* 4(2): 543-573.
- Johnston, R. J., et al. (2017). Contemporary Guidance for Stated Preference Studies. *Journal of the Association of Environmental and Resource Economists*, 4(2), dx.doi.org/10.1086/691697

- Ladenburg, J. and S. B. Olsen (2014). Augmenting short Cheap Talk scripts with a repeated Opt-Out Reminder in Choice Experiment surveys. *Resource and Energy Economics* 37: 39-63.
- Loomis, J. B. (2014). Strategies for overcoming hypothetical bias in stated preference surveys. *Journal of Agricultural and Resource Economics*, 39(1), 34-46.
- Morrison, M., & Brown, T. C. (2009). Testing the Effectiveness of Certainty Scales, Cheap Talk, and Dissonance-Minimization in Reducing Hypothetical Bias in Contingent Valuation Studies. *Environmental and Resource Economics* 44(3), 307-326.
- Norton, D., & Hynes, S. (2014). Valuing the non-market benefits arising from the implementation of the EU Marine Strategy Framework Directive. *Ecosystem Services* 10, 84-96.
- Parsons, G. R. and K. Myers (2016). Fat tails and truncated bids in contingent valuation: An application to an endangered shorebird species. *Ecological Economics* 129: 210-219.
- Train, K. (2016). Mixed logit with a flexible mixing distribution. *Journal of Choice Modelling* 19, 40-53.