Enhancing the Face Validity of Choice Experiments: a Simple Diagnostic Check

Glenk, K; Meyerhoff, Jürgen; Colombo, Sergio; Faccioli, Michela

Published in:
Ecological Economics

DOI:
10.1016/j.ecolecon.2024.108160
10.1016/j.ecolecon.2024.108160

First published: 01/07/2024

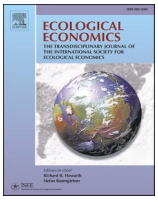
Document Version
Publisher's PDF, also known as Version of record

Ecological Economics 221 (2024) 108160

Contents lists available at ScienceDirect

Ecological Economics

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



Enhancing the face validity of choice experiments: A simple diagnostic check

Klaus Glenk [a,*], Jürgen Meyerhoff [b], Sergio Colombo [c,d], Michela Faccioli [e,f]

[a] SRUC, Department of Rural Economy, Environment & Society, Edinburgh, UK
[b] Department of Business and Economics, Hochschule für Wirtschaft und Recht, Berlin, Germany
[c] Department of Agricultural Economics and Policy, IFAPA, Granada, Spain
[d] WEARE—Water, Environmental and Agricultural Resources Economics Research Group, Universidad de Córdoba, Córdoba, Spain
[e] School of International Studies & Department of Economics and Management, University of Trento, Trento, Italy
[f] Land, Environment, Economics and Policy (LEEP) Institute, Economics Department, Business School, University of Exeter, UK

ARTICLE INFO

JEL codes:
D61
Q510

Keywords:
Validity
Credibility
Stated preference methods
Willingness to pay
Overshooting
Choice experiment design

ABSTRACT

We propose a simple diagnostic check for face validity assessment of willingness to pay (WTP) estimates derived from choice experiments (CEs). The check is based on a threshold value for WTP that is related to the highest cost attribute level, which can be used to assess plausibility of estimated WTP. If the threshold value is exceeded, WTP estimates are considered to overshoot. This may be due to issues with (i) the design of the CE and/or (ii) respondents' behavior deviating from assumptions underpinning CEs. Applying the check to a sample of publications, this paper provides evidence on the incidence and magnitude of overshooting of WTP in the agricultural and environmental CE literature. Based on a random sample of publications including 304 observations representing individual studies and population samples, the results show that overshooting of WTP is widespread, with 65% of observations exceeding the overshooting threshold value. An exploratory analysis to identify factors associated with overshooting of WTP across studies reveals that study design factors, and in particular the design of the cost attribute, play an important role. We recommend that researchers apply the diagnostic check for the design of choice experiments and to motivate further scrutiny of choice experiment results.

1. Introduction

Choice experiments are a stated preference method widely used for non-market (environmental) valuation to inform policy decisions. Choice experiments have rapidly increased in popularity since the early 2000s and is now broadly similar with the contingent valuation method in terms of numbers of publications (Hanley and Czajkowski, 2019). To propagate the use of non-market valuation methods, it is important to investigate and demonstrate their validity and reliability (Bishop and Boyle, 2019).

This paper introduces and discusses a diagnostic face validity check to assess whether willingness to pay (WTP) estimates derived from choice experiments should be considered plausible. This study was motivated by the fact that the authors time and again read or reviewed choice experiment studies that raised dissatisfaction, because the very high WTP estimates reported did not appear plausible given the design of the study. In some of these studies, the reported marginal WTP estimates for changes implied by single attributes appeared to be of reasonable magnitude, but WTP estimated for bundles of attributes was often found to considerably exceed the value of the highest cost attribute level shown to respondents. A typical example would be that a choice experiment study with $100 as the highest cost attribute level reported non-marginal welfare measures of several hundred dollars.[1] We argue that such a study fails to pass a simple face validity test of internal coherence with the study design, and therefore should be subjected to further testing of its validity.

We understand face validity here as a basic yet fundamental form of validity that relates to the degree to which WTP estimates are intuitively

* Corresponding author at: SRUC, Department of Rural Economy, Environment & Society, West Mains Road, EH9 3JG Edinburgh, UK.
E-mail addresses: klaus.glenk@sruc.ac.uk (K. Glenk), Juergen.Meyerhoff@hwr-berlin.de (J. Meyerhoff), sergio.colombo@juntadeandalucia.es (S. Colombo), michela.faccioli-1@unitn.it (M. Faccioli).

[1] We do not provide references here as it is not to our objective to critique single studies in isolation; the numbers chosen should only emphasise commonly found gaps (in relative terms) between highest cost levels included in a study and subsequently presented welfare measures.

https://doi.org/10.1016/j.ecolecon.2024.108160
Received 26 June 2023; Received in revised form 25 February 2024; Accepted 1 March 2024

plausible based on common sense. In this paper, we propose a simple diagnostic check to indicate concern about the overall validity of a choice experiment study. We suggest using this check as a starting point to further scrutinize the plausibility and validity of the estimated WTP measures.

The proposed check applies a criterion for face validity that is defined as follows. Estimates of WTP derived from choice experiments are considered *plausible* within a face validity assessment if sample average estimates of WTP for the non-cost attribute bundle yielding the greatest utility are equal to or lower than a threshold value, which is determined by the highest level of the cost attribute.[2] In choice experiments, the highest cost attribute level represents the maximum payment that respondents would be required to pay if the alternatives presented in the study were implemented in the real world. Given the cost levels displayed in the different alternatives from which respondents make choices, it is to be expected that the average respondents' WTP should not fall too far off the range of values represented by the considered cost attribute levels. Accepting (at face value) WTP estimates that exceed the highest cost level threshold would also imply accepting that respondents' sensitivity to money (marginal utility of income) remains constant outside the range of income change induced by the payments required in the experiment. Whether this assumption is too restrictive may be subject to debate and will be context dependent; from a theoretical perspective, the constant marginal utility of income assumption holds only for goods that make up a small part of consumer spending and decreases for goods that account for a large part of consumer spending (Martin, 2019).

Violations of the criterion set out above imply *overshooting* of WTP and indicate that the estimates derived from CEs are likely biased. This could be broadly explained by the presence of problems (i) in the design of the study, for example regarding the choice of the cost attribute levels (e.g., not high enough), and/or (ii) in the analysis, for example regarding the misspecification of the choice model (i.e., erroneously assuming that respondent behavior abides to the rational choice and utility maximization assumptions).

The proposed criterion is founded in the consideration that implausibly high WTP estimates should raise concerns about the validity of study results, an issue that has been pointed out in a number of choice experiment research contexts (Hess and Beharry-Borg, 2012; Scarpa et al., 2009; Crastes dit Sourd, 2023; Rollins, 2023). In the stated preference literature, face validity has thus far primarily been related to content validity of the survey instrument, which is concerned with "whether the [stated preference] survey asked the right questions in a clear, understandable and appropriate manner with which to obtain a valid [maximum WTP] estimate" (Bateman et al., 2002, p.304). In this paper, we relate face validity to criterion validity, in the sense that preference elicitation through choice experiments should be aligned with external measures that are considered to be 'true', or that are at least closer to the theoretical construct of investigation (Rakotonarivo et al., 2016).

Some additional remarks on the proposed criterion are in order. It is theoretically *possible* that 'true' mean WTP is greater than the threshold value. WTP is at the upper bound constrained by available income, and thus any amount of WTP between zero and that upper bound is theoretically possible. Arguably, however, that does not make any WTP estimate within the interval of zero to available income equally likely. Further, it is often difficult (if not impossible) to empirically establish the deviation of estimated WTP from 'true' WTP, especially in public good contexts, and even when incentivized experiments are used (Colombo et al., 2022). Given the above, it is therefore not advisable to blindly accept any WTP estimate that falls within the possible interval range of values between zero and available income. In a hypothetical example, imagine a choice experiment valuing chicken breast meat with organic and animal welfare credence attributes. Following common practice, price levels are informed by market prices and the highest price level included in any choice alternative for a pound of breast meat is $10. After data is collected and analyzed, mean WTP for a pound of organic and animal welfare friendly chicken breast meat is estimated to be $30. While it is possible that some respondents are willing to pay more than $10, and $30 clearly falls within the interval bound only by available income, it is unlikely that the 'true' WTP of the average respondent is three times higher than the highest market price for the good. One can also question whether results that exceed any price at which a hypothetical "purchase" was made in the choice experiment are perceived as valid by respondents, researchers, or policy makers. We suggest that studies with empirical findings that violate our proposed criterion should be further assessed to understand the possible causes of WTP overshooting.

Based on the above, the main objective of this paper is to investigate the incidence and magnitude of overshooting of WTP in the agricultural and environmental economics literature. To gauge the level of concern regarding overshooting of WTP, we initiated a systematic inquiry of the environmental and agricultural economics choice experiment literature, which has grown significantly over the past three decades (Hanley and Czajkowski, 2019). Common with a systematic literature review, we use a method to summarize evidence (see Section 3). However, in contrast to systematic reviews we do not aim for inclusion of all relevant studies (i.e., the entire agricultural and environmental choice experiment literature) but draw on a random sample of the relevant literature (for the period 2000–2018) to gather evidence on the incidence and magnitude of overshooting of WTP.

To enable an objective and fair comparison across different studies, we apply a simple criterion for evaluating whether overshooting of WTP is present in a study or not. The simplicity of the criterion is not only of practical advantage (easy to identify and implement), but also allows us to draw on a large pool of studies. Such a large database is advantageous for understanding the extent of the problem and to explore the role of possible factors that may contribute to passing or violating the criterion. Importantly, the simple criterion also allows a quick and transparent check for researchers and policy makers to decide if welfare estimates require further assessment. The use of the face validity check may also be relevant when analyzing pilot samples to test the adequacy of the study design or when deciding whether and how to use the estimated WTP to inform policy.

While face validity has been criticized for being a "loose concept", it has also been praised as a way to establish whether data is "fit for purpose" (OECD, 2017, p60). We therefore believe that our proposed diagnostic check, while possibly an over-simplification, can play a role in driving forward a much needed "credibility revolution" in stated preference research, as highlighted by Wiktor Adamowicz in his keynote speech at the International Choice Modelling Conference 2022. This revolution is above all expected to be driven by an increasing role of incentive compatibility and consequentiality that can help enhance the validity of choice experiments. To strengthen the validity of stated preference methods, sound experimentation is needed to allow inference of causal effects. Our simple check cannot and should not replace the numerous studies dedicated to formally testing validity of choice experiments. Rather, it may serve as a quick first assessment of the overall validity of studies and may inspire greater scrutiny of study design and findings to generate, broadly speaking, more plausible WTP estimates for inclusion in policy and decision-making.

As a short preview of results, our analysis indicates that overshooting of WTP is a widespread phenomenon in the agricultural and environmental choice experiment literature. In addition, we find that the design of the cost vector is a main driver of overshooting of WTP. The cost vector is a central element in any stated choice experiment to estimate the welfare measures, but has thus far received relatively little attention in the literature (Glenk et al., 2019; Ahtiainen et al., 2023).

[2] The threshold value is not equal to the highest cost attribute level for all cases, as we explain in Section 3.2.

This paper proceeds as follows. The next section will provide an overview of potential factors contributing to overshooting of WTP (Section 2). Section 3 is a description of the process to identify and select the random sample of choice experiment studies, and will introduce the approach taken to analyze the data. Section 4 reports the results of this analysis, followed by an exploratory analysis of potential aspects that may contribute to overshooting of WTP. Section 5 will discuss results, provide recommendations for choice experiment research and application, followed by conclusions (Section 6).

2. Factors contributing to overshooting of WTP

The discussion of factors promoting WTP overshooting can be structured around two overarching reasons: first, design effects; second, insufficient or ill-reflected representation of respondent behavior when modelling discrete choices. Both groups of factors can also be related if, for example, information provided to study participants as a design feature affects respondents' behavior. Given the context dependency of choice experiment results established in numerous studies (for an overview see Faccioli and Glenk, 2022; Mariel et al., 2021), this implies a potentially long list of sources of overshooting, and options for its mitigation, that cannot be comprehensively discussed in this paper. We therefore focus on those potential sources that have been linked with low measures of sensitivity to cost (i.e., low magnitude of estimates of marginal utility of income), which result in very high WTP estimates.

Concerning the design effects, the most significant driver of overshooting in WTP might be related to the selection of the cost attribute levels (the cost vector). While there is limited guidance on the choice of the cost vector in stated preference studies, we can nevertheless draw on lessons from a small amount of studies investigating the role of cost vector choice on WTP estimates. First, the highest level of the cost attribute should ideally be chosen to be sufficiently high to choke off demand; otherwise, measures of marginal utility of income may be low in magnitude and associated WTP estimates may be inflated (Mørkbak et al., 2010; Glenk et al., 2019; Ahtiainen et al., 2023). The findings from plotting of 'bid acceptance curves', which show the relationship between percentage of alternatives chosen by cost attribute level, suggest that it is a challenging task to completely choke off demand: a residual number of respondents appears to continue to choose alternatives even if magnitudes of cost increase considerably (Kragt, 2013; Mørkbak et al., 2010; Glenk et al., 2019). This is akin to the problem of 'fat tails' in contingent valuation (CVM) studies, as discussed in Parsons and Myers (2016), and in line with results from previous studies that tested the effect of the choke price on respondent' WTP in CVM (e.g., Kanninen, 1995). Continued demand despite a very high cost of alternatives may be legitimate if respondents have high levels of wealth. However, this should plausibly only apply to a relatively small proportion of study participants. Other reasons for a high degree of 'residual bid acceptance' fall into the behavioral domain and include aspects that may be indicative of the presence of hypothetical bias including cut-off violations (e.g., Colombo et al., 2016), anchoring (e.g., Chien et al., 2005), yea-saying (e.g., Brown et al., 1996) and 'non-attendance' to cost (e.g., Scarpa et al., 2009). Earlier findings from CVM studies (e.g., Kanninen, 1995) suggest that placing bids in the extreme tails of the WTP distribution should be avoided, and Mørkbak et al.'s (2010) investigation of choke price bias in choice experiments confirms that choosing the highest cost level beyond a point where bid acceptance decreases only marginally with increasing cost will result in a significant increase in WTP estimates.

The choice of the lowest level of the cost vector has also been shown to affect WTP estimates. This is described in detail in the Appendix of Hess and Beharry-Borg (2012). The argument is that the utility associated with the difference between the lowest level of the cost vector and the status quo option, which often takes a value of zero, is captured by the constant (typically either associated with the status quo alternative or the policy alternatives). This is not a problem if marginal disutility of parting with money is constant over the whole range of cost offered in the choice experiment, i.e., from zero to the highest cost level. However, if marginal disutility is larger in magnitude when the payment required for the hypothetical choice alternative is low, compared to high (i.e. if cost sensitivity is non-linear), a greater difference between zero and the value of the lowest cost attribute level results in cost coefficients that are smaller in magnitude, compared to a situation where the cost coefficient was estimated over the whole cost range (i.e., from zero to the maximum cost attribute level). This, in turn, results in potentially much higher estimates of (marginal) WTP. Appendix Fig. A1 shows a graphical illustration of the above. Apart from evidence provided in Hess and Beharry-Borg (2012) and a discussion of the issue in Villanueva et al. (2017) in the context of willingness to accept (WTA) for agri-environmental contracts, we are not aware of any systematic attempts to assess the severity of the issue in the literature. This applies in general to potential bias associated with non-linear cost sensitivity over the range of cost observed.

We now turn to behavioral factors affecting overshooting of WTP. Obviously, all factors contributing to hypothetical bias are potential determinants of overshooting of WTP. This includes anchoring, yea-saying, and forms of strategic behavior. The use of *ex ante* mitigation devices for hypothetical bias such as cheap talk (Cummings and Taylor, 1999; Carlsson et al., 2005; Tonsor and Shupp, 2011), repeated opt-out reminders (Ladenburg and Olsen, 2014) or other forms of *ex ante* mechanisms such as honesty priming (de Magistris et al., 2013; Howard et al., 2017) or oath scripts (Carlsson et al., 2013) can likely affect the incidence and magnitude of overshooting, with the expectation that use of hypothetical bias mitigation devices decreases the likelihood and magnitude of overshooting. This also applies to *ex post* calibration of WTP based on certainty scales (Brouwer et al., 2010) or *ex post* elicitation of WTP with subsequent opportunity for choice revision (Colombo et al., 2016), or to a combination of *ex ante* and *ex post* approaches (Colombo et al., 2022).

Overshooting may also be affected by respondents' perceptions that the choice experiment will have some real-world impacts and will have actual consequences in the form of a payment for them. An increasing number of studies investigate the role of consequentiality and incentive compatibility in choice experiments (Czajkowski et al., 2017), often in the form of perceived consequentiality (e.g., Petrolia et al., 2014) or policy and payment consequentiality (Zawojska et al., 2019). Consequentiality has been found to enhance truthful value elicitation (Vossler et al., 2012). Therefore, a greater degree of perceived consequentiality may be expected to result in lower WTP estimates. However, to our knowledge, the predominant and somewhat paradoxical empirical finding is that a greater degree of perceived consequentiality is associated with *higher* WTP estimates (Petrolia et al., 2014; Welling et al., 2022). Therefore, perceived consequentiality may affect overshooting of WTP, but more research is needed to understand the empirical findings regarding perceived consequentiality.

Among other behavioral factors, the use of simplifying information processing strategies can be another source of WTP overshooting. This applies especially to 'non-attendance' to cost. Some respondents are found to ignore the cost attribute when choosing among choice alternatives, resulting in low mean sensitivities for the cost attribute and consequently inflated WTP estimates (e.g., Scarpa et al., 2009; Glenk et al., 2015; Rollins, 2023). It is, however, difficult to empirically isolate the potential reasons for inattention to the cost attribute (Alemu et al., 2013), and some authors argue that what is established empirically as non-attendance to cost may rather be low sensitivity to cost (Hess et al.,

2013).

Another behavioral aspect that is potentially related to overshooting of WTP, and that requires consideration at the experimental design stage, is the presence of substitution patterns between attributes. Such patterns are generally modelled through the inclusion of two-way attribute interactions (Riera et al., 2012). Ignoring substitution patterns when they are present may result in higher welfare estimates for attribute bundles yielding the greatest utility (Schaafsma and Brouwer, 2020).

Fundamental questions may also be raised about choice experiment respondents' perception and strategic behavior related to the 'market mechanism' that choice experiments aim to mimic. Respondents may accept alternatives at a cost that is higher than their actual WTP to appear as 'responsible citizens', who want to signal governments to act on an environmental problem. This may result in *act utility*, i.e., the utility associated with endorsing a desirable outcome or with being responsible for the realization of the outcome (Comerford and Hanley, 2017). A manifestation of act utility would be warm glow (Andreoni, 1990).

The above discussion illustrates the breadth of potential mechanisms that could contribute to overshooting of WTP in choice experiment studies, underlining the importance of inspecting the possible contribution of aspects related to study design or behavioral factors to provide recommendations on how to reduce their impact. However, it must be kept in mind that some mechanisms are related to each other and may simultaneously be at work in a study that exhibits overshooting. This highlights that it may be impossible to make a generally applicable statement about the relative importance of factors explaining the incidence and magnitude of overshooting of WTP. This especially applies if the investigation covers a very broad range of applications across the agricultural and environmental economics domains, as is the case in this paper.

3. Methods

3.1. Sample

To investigate the incidence of overshooting of WTP within published choice experiment studies, a random sample of articles was drawn using a modified PRISMA approach (Moher et al., 2009) that relied on the Web of Science database to identify choice experiment papers published in the list of journals reported in Appendix Table A1. The search path applied was "PY = 2000–2018; TS="choice experiment" *OR* TS="choice modelling" *OR* TS="choice modelling"" (PY = Publication Year; TS = Topic). This yielded 642 publications, from which we randomly drew a sample of 301 papers (see Supplementary Materials for a complete list).[3]

We restricted our analysis to studies with a clear (environmental) public good demand context, or a private good context with clear environmental management relevance. This includes studies on outdoor recreation (unless distance is used as proxy for cost),[4] and choice experiments investigating demand for water, fuel, energy and food consumption (which are private goods with public goods repercussions). After screening the papers for suitability to be included in the analysis, 114 papers were dropped for reasons detailed below. 25 papers employed valuation approaches that were not considered here, such as experimental auctions, benefit transfer, best-worst scaling or contingent ranking; and six papers were reviews of the literature or presented no empirical application of a choice experiment. In 16 cases, papers presented choice experiment studies in WTA format, mostly related to public good supply, for example investigating farmers' WTA to participate in agri-environmental schemes. The main reason for dropping a paper, applying to more than half of the sample, was missing or incomplete information to allow WTP estimation and derivation of thresholds for overshooting of WTP. We expended considerable time and effort to obtain any missing information from working papers or earlier versions of the papers, including by contacting the authors of those studies. We were not always successful in filling the gaps and had to exclude 37 papers with missing or incomplete general information regarding the cost vector or the non-cost attributes. In other cases, papers were dropped despite reporting on choice experiments. However, they were not designed to estimate WTP, or because it was not possible to straightforwardly estimate marginal WTP for attributes from the available information. Estimation of marginal WTP was challenging if, for example, a study modelled numerous interaction terms with attributes, often coinciding with missing information on means of interacted variables; or if a study entailed complex functional forms of utility functions that did not allow deriving marginal WTP estimates for undertaking the simple face validity check (e.g., non-linear functional forms where marginal WTP depends on the level that an attribute takes). A small number of papers were excluded, because they do not relate in any way to environmental contexts (four cases) or because they are difficult to access due to language barriers (five cases). Finally, 13 papers were removed because the dataset had been used for another paper in the sample. In this case, as a convention we retained the paper that was published most recently using the same dataset as earlier papers.

The final database used for analysis comprises of 187 papers, which yield 304 independent studies or observations. Each observation refers to model results of data collected from an independent sample of respondents and, importantly, does not refer to results of different model types or specifications based on the same sample. If more than one model type or specification was reported for a given sample, we used results of the model that was indicated to be preferred by the authors of the study. In the absence of such information, we selected the best model mostly by relying on model fit criteria. Information on such decisions and their rationale were systematically recorded to ensure transparency. The final database, including comments about assumptions and models used for further assessment, is available as Supplementary Material.

3.2. Data extraction

The main aim of this study is to investigate how prevalent overshooting of WTP is among choice experiment studies applied to agricultural, food or environmental context, using a diagnostic check to indicate if WTP estimates overshoot. This requires information on: (i) mean marginal WTP associated with the attributes to calculate sample average estimates of WTP for the non-cost attribute bundle yielding the greatest utility, and (ii) the threshold value based on cost attribute levels that determines whether WTP as identified in (i) is considered to overshoot or not, depending on whether it crosses the threshold value. A summary guide (with examples) providing an overview of the main steps of the diagnostic check is provided in the Appendix, Box A1.

Estimates of mean marginal WTP for attribute levels are extracted from the source publications, if reported. In a first step, we identified the attributes and attribute levels that jointly yield the alternative with the greatest possible utility based on model results. Then, we recorded marginal WTP estimates for these attributes and attribute levels, based on the estimates provided in the publications. If WTP estimates were not

[3] We have undertaken a detailed *post-hoc* sample size and power analysis to explore whether sample size is sufficient for the inference of WTP overshooting incidence and relative magnitude. We approached the analysis by investigating the scale of benefits gained in terms of precision (reduction of margin of error) of proportion and relative magnitude of studies failing the face validity check, against the significant cost implied in terms of data entry. We find that the benefits of additional data entry are relatively low, while cost in terms of time and effort is considerable.

[4] Given the threshold criterion formulated in this study to identify overshooting relates to cost attribute levels, only studies including a cost attribute could be considered in the analysis. This represents the approach followed by the vast majority of choice experiment studies.

available from the publications, we used the reported model results to calculate the mean marginal WTP ourselves (i.e., by dividing the mean of the parameters of the non-cost attributes divided by the mean of the cost attribute). If these calculations were necessary, we included a comment in our database for transparency.

To determine the threshold that defines if a study exhibits overshooting of WTP or not, we first recorded the levels of the cost attribute (the cost vector). If the cost vector entails negative values, the overshooting threshold is defined by the difference between the maximum cost level and zero since we are interested in WTP. If the cost vector has only non-negative values, then the overshooting threshold is established as the difference between the maximum and the minimum cost attribute level. Note that zero is often not part of the cost vector entering the experimental design but used as the cost of the reference alternative (e.g., the status quo). Therefore, in many studies the minimum cost attribute level will be the lowest non-zero (positive) level of the cost attribute.

It may seem intuitive to simply use the maximum cost attribute level as the threshold. This would work reasonably well for many studies, especially those with an environmental public good focus, where the status quo alternative is offered at no extra cost to respondents (most studies in our sample). However, it can be problematic to solely rely on the maximum cost attribute level for studies that have a non-zero (positive) price for the reference or status quo alternative. For example, in the food choice domain, it is common to define the reference alternative based on the respondents' usual purchase price of a product (generally non-zero). The reference price of the reference product and the cost attribute levels of the product alternatives may be defined in absolute terms (for example $5 for a reference chicken breast pack, and attribute levels in the remaining alternatives as $6, $7 etc.); however, the reference price may not be identified, and cost attribute levels may be defined relative to the reference product (for example, as a price increase of $1, $2 over the usually purchased product). Relying on the maximum cost attribute level to define the overshooting threshold would in such cases imply different thresholds for different respondents. Another argument for not using the maximum level of the cost attribute, but the difference between the maximum cost attribute level and the lowest non-zero (positive) level, is that this represents the range of values over which marginal disutility of cost is derived as the basis for WTP estimation.

Note that our definition of the threshold for overshooting does not include utility captured via alternative specific constants (ASCs). In many contexts, it is important to consider the ASC for deriving welfare measures. However, we investigate face validity of choice experiments through a simple criterion that indicates whether WTP estimates are plausible. In many choice experiment applications, ASCs capture utility differences between zero and the lowest cost and possibly non-cost attribute levels (Hess and Beharry-Borg, 2012; Villanueva et al., 2017). They also capture systematic tendencies to select status quo, reference or opt out alternatives that are not explained by differences in attribute levels across the alternatives. Such behavioral tendencies are arguably relevant for welfare estimation. However, they offer little to an evaluation of whether the design and model of a choice experiment application *enables* plausible WTP estimates. Also, ASCs associated with the status quo alternative may be negative and therefore suggest a preference for policy change. Further, if the ASC indicates a tendency to choose the status quo alternative (status quo bias), this is typically at least to some degree the result of serial non-participants whose WTP can genuinely be assumed to be zero. In this case, marginal WTP for the attributes should still be identified in a plausible manner for the remainder of the respondents. The same applies to cases where a status quo bias is found for participants.

We also extracted additional information from the sampled papers to characterize the studies and to conduct an exploratory regression analysis to understand which factors may be associated with a greater or smaller incidence and magnitude of WTP overshooting. As we will discuss, uncovering the reasons for overshooting is challenging, especially given a sample of studies covering a wide range of topics. Nevertheless, our exploratory analysis can be seen as a starting point for wider discussion and in-depth analysis going forward.

Extracting information from publications is not always straightforward in the absence of reporting conventions, and in the presence of methodological changes over time. When selecting variables to record for each observation, we therefore focused on aspects that can be unambiguously identified. For example, we recorded information on the number of attributes and alternatives and overall design size, but the final database does not entail details on the type of experimental design used (orthogonal; D-optimal; D-efficient etc.) since the reporting practice in publications did not allow a meaningful and clear categorisation in a considerable number of cases. All four co-authors added studies to the database. To ensure coherence of information across researchers, each entry was first checked visually by the corresponding author, and any issues were clarified with the researcher who had entered the data based on a joint reading of the publication. Furthermore, each co-author cross-checked in detail a random selection of 30 studies that had been entered by a different co-author. No major disparities emerged as a result of these in-depth spot checks. The only changes applied to the database after these checks consisted in the addition of further details to existing comments regarding individual entries, with the goal of enhancing transparency of reporting.

3.3. Indicators of overshooting and statistical analysis

For the purposes of our research, we identified two main variables of interest from our constructed database of studies. First, whether overshooting occurred, i.e., if total WTP for the bundle of attributes yielding the highest utility was greater than the threshold value. Second, we calculate a measure of relative deviation from the threshold value for overshooting of WTP, i.e., the percentage by which WTP overshoots and exceeds the threshold. These two variables also serve as dependent variables in a binary logistic regression (overshooting yes/no) and in an ordinary least squares (OLS) regression (semi-log; natural log of % overshooting) with a subset of variables listed in Table 1 as regressors. The logit model investigates which factors can affect a study's likelihood to show overshooting, while the OLS investigates which factors influence the relative magnitude of overshooting for those studies that overshoot. For both types of regression, we employed a "manual backward selection" approach. In the first step, all variables reported in Table 1 are entering the models. After the initial run, the variable with the highest *p*-value was dropped from the model if it was above $p = 0.2$. In case of dummy groups (e.g., location of study or survey method, see Table 1), the whole group remained in the model if one of the dummies belonging to this group was $p = 0.2$ or below. Subsequently, the models were re-run until no further variables could be excluded given the above mentioned threshold. To account for the fact that multiple observations could stem from the same publication, clustered errors for all estimates derived from the same study were calculated. When estimating the OLS regressions, we also applied outlier analysis using residual versus fitted value plots and leverage versus squared residual plots (Cameron and Trivedi, 2010).

4. Results

4.1. Overview of the sampled studies

Reflecting the surge in choice experiment applications in recent years, about half of the observations are from studies published after 2012 (Table 1). In terms of regional distribution, studies from Europe provide approximately half of the observations, followed by New Zealand and Australia (18% of the observations), and North America (14% of the observations). 75% of studies covered an (environmental) public good context, with the remaining 25% focusing on private goods. For the

Table 1
Summary statistics for selected variables extracted from the sampled publications.

Variable name	Description	Mean	Std. Dev.	Min	Max
Year of publication					
til2008 (*)	Study published before 2009†	0.22	0.42	0	1
til2012	Study published after 2008 but before 2013†	0.33	0.47	0	1
til2018	Study published after 2012 but before 2019†	0.45	0.50	0	1
Location of study					
Europe (*)	Case study: Europe†	0.48	0.50	0	1
NorthAmeric	Case Study: North America†	0.14	0.35	0	1
NZAus	Case Study: New Zealand or Australia†	0.18	0.39	0	1
CountryOther	Case Study: Other region†	0.17	0.38	0	1
Payment vehicle (PV)					
PV_M (*)	Payment vehicle: mandatory†	0.84	0.37	0	1
PV_V	Payment vehicle: voluntary†	0.06	0.24	0	1
PV_G	Payment vehicle: general/ unspecified†	0.11	0.31	0	1
Model used for WTP estimation					
M_CL	Model: conditional logit†	0.22	0.41	0	1
M_RPL	Model: random parameter logit (RPL) †	0.58	0.50	0	1
M_RPL_fix‡	Model: RPL with fixed cost†	0.39	0.49	0	1
M_RPL_ran‡	Model: RPL with random cost†	0.17	0.37	0	1
M_LCM	Model: Latent class†	0.05	0.22	0	1
M_EC	Model: Error component logit†	0.05	0.22	0	1
M_NL	Model: Nested logit†	0.06	0.23	0	1
M_Other	Model: Other (e.g., Hierarchical Bayes Random Parameter Logit, Multinomial Probit, Heteroskedastic Logit)†	0.07	0.26	0	1
Survey mode (SM)					
SM_mail (*)	Survey mode: mail†	0.26	0.44	0	1
SM_online	Survey mode: online†	0.25	0.43	0	1
SM_face2	Survey mode: face to face†	0.41	0.49	0	1
SM_other	Survey mode: Other/mixed†	0.08	0.28	0	1
Cost vector characteristics					
Count_price	Number of levels in cost vector	5.62	1.66	2	22
Minspread	(Maximum cost level – Minimum cost level) / Minimum cost level	10.95	15.05	0.10	132.33
Other design features					
N_alt	Number of Alternatives	3.09	0.68	2	7
N_att	Number of Attributes	5.09	1.73	2	13
N_set	Number of Choice Tasks	7.01	3.27	1	26
CheapTalk	Study mentions use of cheap talk script†	0.11	0.31	0	1
Other study characteristics					
C_private	Context: private good (PG) focus†	0.25	0.44	0	1
MethodFocus	Paper focus on methodological advances†	0.48	0.50	0	1

Note: (*) reference value in the regression models; † Dummy variable (yes = 1, else 0); based on 304 studies extracted from 197 publications; ‡ reported for completeness, since not significant in regressions and thus not statistically different from effects of M_RPL.

applications concerning environmental public good provision, there is a wide and relatively even distribution of topics covered. One third of studies focusing on private goods are related to food choices, followed by applications concerning fuel and energy consumption (26%) and recreation (22%).

Approximately half of the sampled papers are deemed to have a dedicated focus on methodological advances as the main objective of the research, based on the judgment of the authors of this paper. However, the authors of the source studies may disagree with our judgment. We acknowledge that, in some cases, studies that are not classified as having a methodological focus may indeed have addressed some methodological issues as well. Irrespective of the degree of subjectivity involved in the classification, our sample has good representation of publications with both an applied and a methodological focus.

The payment vehicle used in most of the choice experiments included in our database is predominantly mandatory (tax, change in water or electricity bills). We find a significant variation in the number of cost attribute levels forming the cost vector of the reviewed studies, with an average of between 5 and 6 levels. The average choice experiment study in our sample has 3 alternatives, 5 attributes and 7 choice tasks, with modest degrees of variation in each of these design dimensions. Observations based on face-to-face surveys are most frequent (41%), followed by mail and online surveys, which make up approximately one quarter each of the total observations. About 10% of the sampled studies reported to have used a cheap talk script or an opt-out reminder (*CheapTalk*). Implementation of the scripts, however, varied considerably across studies, both in terms of content and length.

Just less than a quarter of the choice experiment studies in the database used conditional logit (CL) models as the only or the preferred model. The predominant model used (58%) is the random parameter logit (RPL) model, where the utility of one or several attributes is allowed to vary across respondents. Among the RPL models, the cost attribute was allowed to vary following a specified distribution in more than two thirds of cases while it was assumed to be fixed in the remainder of observations. Latent Class, nested logit and error component (EC) models (that introduce a common random error term shared by two or more of the choice alternatives) make up smaller shares (around 5% each) of the observations.

Finally, the variable *Minspread* captures the range of values of the cost vector. Greater values indicate a greater difference between the lowest cost attribute level and the highest cost attribute level. For example, the average value of 11 indicates that, on average, the highest cost attribute level was 11 times greater than the lowest non-zero positive cost level. The average number of cost attribute levels (*Count_price*) is 5.62.

4.2. Indicators of overshooting of WTP

Of the 304 observations, overshooting of WTP occurred in 65% of the cases (Table 2). Fig. 1 shows relative deviations of WTP calculated for the attribute bundle yielding the greatest utility from the overshooting threshold. For studies that exhibit overshooting of WTP, the mean relative magnitude of overshooting is 254%. The relative magnitude of overshooting is equal to or below 50% for about a third (30%) of the observations where overshooting of WTP occurs, and half of the observations (50%) displays a degree of overshooting of 100% or less. WTP for the 'best' attribute bundle exceeds the overshooting threshold by more than 200% in a considerable number of cases (37% of the observations where overshooting is present), and in 17% of the sampled studies the relative magnitude of overshooting is greater than 500%. Ten observations are estimated to have a WTP for the 'best' possible attribute level combination that is equal to or greater than 1000% (i.e., ten times) of the overshooting threshold. The results show that the incidence of overshooting remains high even if the criterion for overshooting was relaxed, that is, if the threshold value for overshooting was increased by a certain percentage.

Table 2
Variables indicating the incidence and magnitude of overshooting of willingness to pay (WTP)[†].

Variable	Description	Number of studies in database	Mean	Median	Std. Dev.	Min	Max
OS	Overshooting of WTP (dummy variable)	304	0.65	1	0.48	0	1
OS – private goods	Overshooting of WTP if study focus is on private good (*OS* if *C_private* = 1)	77	0.73	1	0.45	0	1
OS – public goods	Overshooting of WTP if study focus is on public goods (*OS if C_private* = 0)	227	0.62	1	0.49	0	1
OS%	Deviation from the threshold WTP for the attribute bundle yielding the greatest utility (% of threshold value)	304	151.42	37	315.44	−99.98	1859.92
OS% –overshooting studies	Relative magnitude of overshooting if it occurs (*OS%* if *OS* = 1)	197	254.46	100.5	351.02	0.11	1859.92
OS% –overshooting studies – private goods	Relative magnitude of overshooting if it occurs in private good studies (*OS%* if *OS* = 1 & *C_private* = 1)	56	300.40	107.94	373.54	0.67	1324.09
OS% –overshooting studies – public goods	Relative magnitude of overshooting if it occurs in public good studies (*OS%* if *OS* = 1 & *C_private* = 0)	141	236.22	97.44	341.33	0.11	1859.92

[†] WTP values that underpin figures reported in Table 2, and serve as a basis for derivation of dependent variables for regressions reported in Table 3, have been calculated from the database of studies ($n = 304$) for attribute bundles yielding the greatest utility, determined through information provided in sampled publications ($n = 197$).

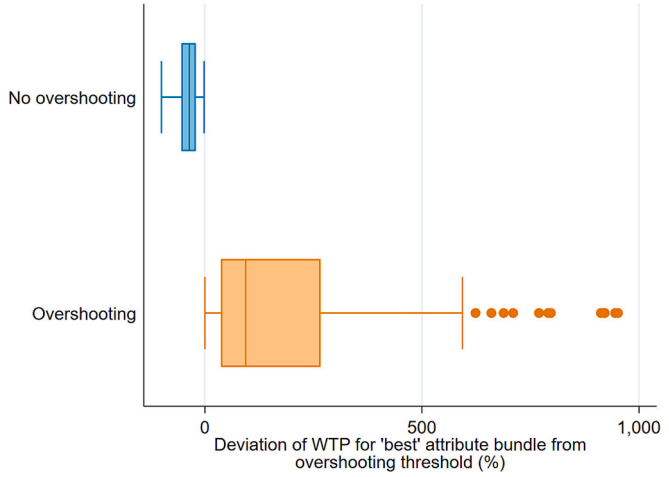


Fig. 1. Box-and-Whisker plots of % deviation of willingness to pay (WTP) for attribute bundle yielding the greatest utility relative to the overshooting threshold. Values truncated at 1000% (293 studies).

Overshooting of WTP is slightly more frequent among the 78 observations with a private good focus (73%) than for those studies that used choice experiments to elicit WTP in an environmental public good context (62%). The magnitude of overshooting of WTP relative to the overshooting threshold is also slightly higher in studies with a private good focus. In studies classified to focus on methodological advances, overshooting of WTP occurs in 45% of cases. In the remaining applied studies, often focusing on the use of valuation to inform a policy development, the incidence of overshooting is slightly higher (55%).

4.3. Exploratory analysis of factors associated with overshooting

The effect of selected factors on overshooting is investigated through binary logit and OLS regressions (Table 3). We present both model results with and without clustered standard errors.

To analyze the factors affecting the likelihood of overshooting, we focus on the upper part of Table 3, reporting the results of the binary logit models. Based on these findings, only few factors show a statistically significant association with overshooting based on the manual backward selection approach (*p*-value equal or lower 0.2). All else equal, overshooting is more likely if the number of alternatives within a choice set increases. In contrast, face-to-face interviews decreases the likelihood of overshooting compared to studies that used mail surveys. A greater spread of the cost vector (*Minspread*; the relative difference between the lowest and the highest cost attribute level) significantly decreases the likelihood of overshooting being present. Also, a greater number of levels of the cost vector (*Count_price*) reduces the likelihood of overshooting. However, in the models reported (with clustered standard errors), this variable is only significant at the 10% level. The variable *Count_price* is positively correlated with a greater relative spread of the cost vector (Pearson's correlation coefficient: 0.34, p-value: 0.00) and therefore possibly captures part of the effect of *Minspread* on the dependent variable. Finally, if the study was recorded to have a methodological focus (*MethodFocus*), overshooting is less likely. This effect, similar to the number of cost attribute levels, is only significant at the 10% level.

Table 3
Results of binary logit model of overshooting and OLS regression of relative magnitude of overshooting.

	Coefficient	Robust standard errors	|z-value|	p-value
Binary logit[†] ($n = 304$)				
M_RPL	−0.440	0.410	1.07	0.283
M_LCM	0.077	0.688	0.11	0.911
M_EC	−0.263	0.651	0.40	0.686
M_NL	−0.994	0.709	1.40	0.161
M_Other	−1.071	0.634	1.69	0.091
SM_online	−0.311	0.428	0.73	0.468
SM_face2	−1.082	0.474	2.28	0.022
SM_Other	0.312	0.984	0.32	0.751
NAlt	0.820	0.306	2.68	0.007
Count_price	−0.181	0.102	1.78	0.075
Minspread	−0.058	0.026	2.26	0.024
MethodFocus	−0.699	0.393	1.78	0.075
Constant	1.066	1.030	1.04	0.301
OLS regression[‡] ($n = 195$[§])				
NorthAmerica	0.024	0.188	0.13	0.898
NZAus	−0.066	0.172	0.39	0.699
CountryOther	−0.278	0.189	1.47	0.144
M_RPL	−0.143	0.158	0.9	0.368
M_LCM	−0.288	0.297	0.97	0.334
M_EC	−0.377	0.210	1.79	0.075
M_NL	−0.182	0.262	0.69	0.489
M_Other	−0.579	0.253	2.29	0.024
SM_online	0.075	0.160	0.47	0.640
SM_face2	0.467	0.174	2.69	0.008
SM_Other	0.144	0.193	0.75	0.457
NAlt	0.131	0.069	1.90	0.059
NAtt	0.071	0.041	1.74	0.084
Count_price	−0.123	0.048	2.57	0.011
Minspread	−0.017	0.009	1.88	0.062
Constant	5.535	0.445	12.43	0.000

Note: [†] Dependent variable: OS (see Table 2); [‡] Dependent variable: OS% (see Table 2); [§] two studies excluded based on outlier analysis; pseudo R-squared logit: 0.16; R-squared OLS: 0.30.

Turning to the analysis of the factors affecting the degree of overshooting (measured as the percentage that WTP estimates exceed the threshold value), we find a significant effect (at the 10% level) for five of the included variables (lower part of Table 3). The number of attributes is found to have a significant effect and more attributes in a choice experiment appear to increase the degree of overshooting. Face-to-face surveys compared to mail surveys, if they overshoot, are also associated with greater relative magnitude of overshooting of WTP. We find that a greater spread of the cost vector (*Minspread*) and a greater number of levels of the cost vector decrease the relative magnitude of overshooting. Studies using uncommon model types including, for example, hierarchical Bayes random parameter logit or heteroskedastic logit models, show significantly lower relative magnitude of overshooting relative to studies whose WTP estimates are based on conditional logit models. However, this finding needs to be taken with caution as the variable is a summary category for relatively uncommon model types, and it applies to only 22 studies in the sample.

5. Discussion and recommendations

The main finding of this research is that using a simple diagnostic check, overshooting of WTP is present in a considerable share of the investigated discrete choice experiment studies. Half of the studies that were included in the analysis present models that report WTP estimates for the attribute bundle yielding the greatest utility that are more than twice as high in magnitude as a threshold value for overshooting. Overshooting occurs both in studies that are published in policy-oriented journals and journals with a stronger focus on methodological advances. It therefore appears that from a face validity perspective, concerns about the plausibility of benefit estimates derived from choice experiments are justified. Our inquiry covers a wide range of applications and journals, and we are confident that expanding the database beyond the final year in our sample (2018) would not result in a marginalization of the problem, especially when considering that overshooting of WTP was neither addressed in recent recommendations of explicit relevance to choice experiments (Johnston et al., 2017), nor in choice experiment guidance documents (Mariel et al., 2021).

In our analysis, we adopted a cautious and conservative criterion for a face validity assessment, by focusing on mean estimates of WTP and by stipulating that studies whose mean values are–by some relative amount–lower than a threshold should pass the face validity test. For most studies, the threshold value is determined by the difference between the maximum cost attribute level and the minimum (non-zero) cost attribute level. Given the expectation of demand for alternatives approaching zero as 'bids' (cost attribute levels) increase, we may expect that *mean* WTP should be lower than the value of the threshold. This would rather justify tightening the criterion. Given the considerable density of studies that are close to the overshooting threshold (see Fig. 1), adopting a more stringent criterion would lead to a notable increase in the incidence of overshooting.

When presenting the idea of overshooting of WTP and the related criterion, we encountered concerns that using the attribute bundle yielding greatest utility would not be appropriate, because respondents may not have evaluated such a bundle in the actual choice experiment. In other words, the attribute bundle yielding the greatest utility may not be an alternative included in any choice task part based on the experimental design, most likely a fractional factorial of some sort. This argument implies that researchers should only have faith in, and estimate WTP for, alternatives with attribute level combinations that were actually offered to respondents. This is, however, at odds with the basic idea of using (fractional factorial) experimental designs, employed by the majority of choice experiment studies. Fractional factorial designs are used as a way to allow (unbiased) estimation of attribute effects when using a full factorial is not possible, for example due to its size. If we i) assume a fractional factorial design is used; ii) if we consider, in line with neoclassical behavioral assumptions underpinning choice experiments, that respondents have full information and well-formed preferences; and iii) if respondents are assumed that they treat each of a series of choice tasks as independent from each other, then it should not matter for the estimation of WTP if a certain alternative was shown to them or not. Respondents' valuations are absolute, that is, respondents appraise the utility of each alternative based on its own merit, independent of other alternatives. Therefore, whether the utility maximizing alternative has been shown to respondents matters only in case of violation of basic assumptions underpinning choice experiments and their design. This is exactly the argument we make when establishing the overshooting criterion: that violating it should prompt researchers to investigate which assumptions do not hold.

What can be learned from the exploratory analysis of factors associated with overshooting of WTP? Only few aspects are found to be significant determinants of the incidence and magnitude of overshooting. Our regressions include two variables characterizing the cost vector. A negative coefficient related to the relative spread of the cost vector may point towards the relevance of defining the lowest cost vector levels to be close to the cost of the reference alternative, in line with Hess and Beharry-Borg (2012). An alternative explanation would suggest that increasing the highest cost vector level decreases the risk of overshooting. Higher cost may help to choke off demand, but–as Mørkbak et al. (2010) have shown–also carry the risk of choke price bias inflating WTP. Cost vector length is positively but only weakly correlated with spread, and we currently lack a convincing explanation to motivate the significant and negative effect of the number of cost attribute levels. Nevertheless, the consistent findings across our regressions show that the design of the cost vector plays an important role in explaining the incidence and extent of overshooting, which deserves further investigation.

An increase in the number of attributes is found to be associated with a greater degree of overshooting. The literature on the influence of choice experiment design dimensions suggests that the number of attributes affects error variance (randomness of choices) rather than WTP estimates (Caussade et al., 2005; Meyerhoff et al., 2015). It is possible that effects emerge with a greater variation in the number of attributes investigated in these studies. For example, Meyerhoff et al. (2015), in an environmental context, investigate differences between samples receiving between four and seven attributes, while the number of attributes in our dataset varies between three and ten. More important may be that design-within-the-design studies keep the information content entailed in the attributes constant to allow for an experimental assessment of the effects attribute number. In our cross section of studies, an additional attribute will also carry additional information to be evaluated by respondents. It is thus conceivable that different choice behaviors and cognitive processes are evoked by increasing the number of attributes, compared to Caussade et al. (2005) and Meyerhoff et al. (2015). An emerging opportunity to conduct choice experiments with high attribute numbers while keeping choice set complexity and risk of overshooting limited are partial profile experimental designs (Meyerhoff and Oehlmann, 2023). Thus far, this type of experimental design has rarely been used in environmental valuation, but it is deemed to be beneficial for studies that include more than five or six attributes.

The number of alternatives in the choice set is also found to affect the incidence and magnitude of overshooting. Although in theory a binary choice format (one alternative plus status quo or opt out) is considered incentive compatible (Carson and Groves, 2007), less than 3 % of studies in our database of publications employed this format. Three quarters of

the studies used three alternatives, and the maximum number of alternatives in any of the included studies is seven. An increasing number of alternatives may lead to biased estimates due to higher choice task complexity. However, respondents may find it easier to identify the alternative that best reflects their preferences as the number of alternatives increases, resulting in a preference matching effect (DeShazo and Fermo, 2002; Zhang and Adamowicz, 2011). Weng et al. (2020) confirm this in a study comparing split samples with up to four alternatives. They state that preference matching takes place when the number of alternatives increases but that, at the same time, complexity can impede choices. Early findings by Meyerhoff et al. (2017) suggest that WTP can be inflated as the number of alternatives increases. The authors attribute this to a likely increase in the use of a price-quality heuristic with increasing availability of alternatives indicated by a significantly greater likelihood to choose alternatives with higher levels of the cost attribute. The price-quality heuristic has been found in the marketing literature to be of particular importance if there is uncertainty about product quality (Rao and Monroe, 1989). This arguably applies to stated preference contexts of most published studies, at least in the private good domain.

Face-to-face surveys are associated with a decrease in the likelihood of overshooting relative to mail surveys, but also with a decrease in magnitude of overshooting for studies that overshoot. Plausible reasons exist in support of both greater and lower likelihood and magnitude of overshooting for face-to-face surveys relative to mail surveys. Mail surveys are prone to self-selection bias in response. Those interested in a survey topic will have a higher likelihood of responding to the survey and possibly a greater chance to express preferences in favor of proposed changes. This may make mail surveys more likely to overshoot. However, there have been concerns about a greater degree of social desirability bias (SDB) in face-to-face surveys (Lindhjem and Navrud, 2011). SDB arises if respondents adjust (often upwards) their answers to match the interviewer's expectations or to be consistent with social norms (Lopez-Becerra and Alcon, 2021). It is unclear if the above biases play a role in explaining our empirical findings, which are not explained through differences in a time trend in the use of survey modes: *both* mail surveys and face-to-face surveys are significantly and negatively correlated with publication year. We thus leave this question to further research.

The absence of statistically significant effects in our explanatory analysis is noteworthy for some variables. First, there is no significant effect of model type, with the reference being conditional logit models. This might be taken as an indication of the primary role of study design over modelling of behavior to enhance the validity of choice experiment studies. It may also point to known problems with random parameter logit models as the most common alternative to conditional logit models with, for example, long tails of the (lognormal) distribution for cost coefficients that result in inflated WTP estimates (Crastes dit Sourd, 2023). Second, there is no significant time trend in overshooting. Issues with overshooting seem to persist over time, despite considerable advances in the choice experiment methodology since the early 2000s. However, there is still little progress regarding the development of guidance on cost attribute design, despite its effect on overshooting, as shown in our explanatory analysis. Third, no significant effect for *ex ante* mitigation strategies aimed at reducing hypothetical bias has been observed. This is somewhat surprising, as the numerous instruments available to counteract this bias in the choice experiment literature (single or multiple opt-out reminders, cheap talk scripts, solemn oath, etc.) should be expected to reduce overshooting. Empirical evidence regarding the result of *ex ante* mitigation strategies in reducing hypothetical bias are currently mixed (Colombo et al., 2022), thus potentially diluting the effect on overshooting. A possible reason contributing to our finding may be the large variation in the implementation of the mechanisms to minimize hypothetical bias in the choice experiment literature. Finally, we do not find a significant difference in overshooting between studies with public and private good context. Arguably, respondents have a greater familiarity with private (market) goods than non-market goods that often have public good characteristics, which justified a significantly greater hypothetical bias found in studies with a public good context compared to private good contexts (Penn and Hu, 2018). Nevertheless, conditions for incentive compatibility of stated preference elicitation for market goods are less clear than for public goods (Entem et al., 2021), and questions about best practice regarding cost vector design remain open for both private and public good studies. Also, within the agricultural and environmental economics literature it might not be possible to clearly assign studies to a private versus a public context, given that private good studies may entail attributes with relevance to public good provision, and vice versa. This intertwined nature of public and private studies may affect the inference of a difference in overshooting by study context.

Recent stated preference guidance (Johnston et al., 2017) emphasizes the importance of using single binary choice tasks as a response format that can, in principle, be incentive compatible (see also Carson and Groves, 2007). It is beyond the scope of this paper to demonstrate if such response formats will be negatively associated with the incidence of overshooting of WTP. Such an inquiry would also benefit from an adjusted focus of investigation on choice-based stated preference methods, thus considering studies previously not labelled as choice experiments but as choice-based contingent valuation.

What could researchers do to reduce the risk of overshooting to occur? Our discussion first focuses on options related to the design of the cost vector. Because of the differences in how WTP estimates are generated using CVM data compared to choice experiments, *ex post* statistical approaches developed to 'pin down' fat tails ('pinched' logit or truncation, see for example Ready and Hu, 1995; Haltia et al., 2009) and thus to derive arguably more plausible WTP estimates are not (yet) available for choice experiments. However, some steps could be taken to mitigate the issue *ex ante*. Based on our results, we believe that greater effort should be expended on analyzing bid acceptance[5] with respect to cost from pre-test and pilot study data with the purpose of better informing the choice of cost vectors. If bid acceptance of the highest cost level is relatively high (e.g., above 10–15%), and the bid acceptance curve does not appear to have flattened out considerably at the highest cost level, it may be advised to increase the magnitude of the highest level of the cost vector. On the other hand, if the bid acceptance curve decreases only marginally between the two highest cost levels and has low values (for example, less than 5%, as illustrated for the choke price bias sample in Mørkbak et al. (2010)), the value of the highest level of the cost attribute should be reduced. The fact that choice experiment surveys are increasingly being implemented online (Liebe et al., 2015) should facilitate this process through the ability to collect pilot study data with relatively large sample sizes. In the past, choice experiment researchers have dedicated effort to improving the efficiency of the experimental design through a sequential Bayesian approach (Scarpa et al., 2007). It would afford little extra effort to combine each round of revision of the experimental design with an assessment of bid acceptance curves as described above. Regarding the impact of the choice of the lowest level of the cost vector, while lacking any further empirical evidence, we would advise against choosing the lowest cost attribute level such that it represents a considerable expense for the average respondent (for example, $50 if the highest cost level is chosen to be $200), in line with arguments outlined in Hess and Beharry-Borg (2012).

It has been a widespread concern that RPL models with lognormally distributed utility for cost tend to generate inflated WTP estimates, especially if estimated in preference-space (e.g., Mariel et al., 2021). With a point mass close to zero, a (negative) log-normal distribution can result in "exploding" and implausible mean WTP estimates, even if median values appear reasonable (Train and Weeks, 2005). While

[5] See Kragt (2013) or Glenk et al. (2019) for detailed explanation on how to estimate and display 'bid acceptance' in choice experiments.

additional regressions did not identify a significant association with overshooting between specifying cost as random or fixed in RPL models, it is nevertheless worthwhile to highlight recent modelling advances that can help to address the issue of implausibly high mean WTP estimates. A recent example is Crastes dit Sourd (2023), who proposes to rely on WTP space-models or, enhancing model performance, to use a shifted log-normal distribution as a novel functional form in RPL models. Rollins (2023) uses mixed latent class-RPL models to account for cost non-attendance which inflated WTP estimates in basic RPL models with log-normally distributed cost. Such modelling advances are important steps towards offering researchers a set of tools that achieves greater face validity of choice experiment estimates.

6. Conclusion

This paper introduces a criterion for a simple and easy-to-implement face validity assessment of WTP estimates derived from choice experiments. For most studies, the criterion uses a threshold value that is determined by the difference between the maximum cost attribute level and the minimum (non-zero) cost attribute level. If the threshold value is exceeded, there are reasons to question the plausibility of estimated WTP. Applying the face validity check to a random sample of 304 choice experiment studies suggests that WTP estimates overshoot relative to the threshold value for a large share of the investigated studies. Characteristics of the cost attribute explain incidence and magnitude of overshooting of WTP. We hence suggest applying the diagnostic check for the design of choice experiments, and to initiate and motivate further scrutiny of results. We believe that its application will support the growing emphasis on credibility of choice experiments to achieve more widespread acceptance of stated preference estimates among decision-makers.

CRediT authorship contribution statement

Klaus Glenk: Writing – review & editing, Writing – original draft, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jürgen Meyerhoff:** Writing – review & editing, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sergio Colombo:** Writing – review & editing, Validation, Methodology, Investigation, Formal analysis, Data curation. **Michela Faccioli:** Writing – review & editing, Validation, Methodology, Investigation, Formal analysis, Data curation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Klaus Glenk reports financial support was provided by Scottish Government Rural and Environment Science and Analytical Services Division.

Data availability

Data is attached as supplementary material

Acknowledgements

KG acknowledges support by the Scottish Government as part of the Rural Affairs, Food and the Environment (RAFE) Strategic Research Programme 2022-2027 SRUC-B3-1: Ensuring positive behavioural change for farmers towards best practice for clean growth: economic and behavioural investigations, and SRUC-D5-1: Understanding the value of Scotland's Agricultural Soil Natural Capital.

Appendix A. Appendix

Table A1
List of Journals included in the database search.

Journals with economics focus	Interdisciplinary journals and journals with policy focus
CANADIAN JOURNAL OF AGRICULTURAL ECONOMICS	JOURNAL OF ENVIRONMENTAL PLANNING AND MANAGEMENT
ENVIRONMENTAL RESOURCE ECONOMICS	JOURNAL OF ENVIRONMENTAL MANAGEMENT
MARINE RESOURCE ECONOMICS	LANDSCAPE AND URBAN PLANNING
JOURNAL OF AGRICULTURAL AND RESOURCE ECONOMICS	ENVIRONMENT AND PLANNING A
AMERICAN JOURNAL OF AGRICULTURAL ECONOMICS	ECOSYSTEM SERVICES
JOURNAL OF AGRICULTURAL ECONOMICS	ENERGY POLICY
AUSTRALIAN J OF AGRICULTURAL AND RESOURCE ECONOMICS	MARINE POLICY
LAND ECONOMICS	TOURISM MANAGEMENT
ENERGY ECONOMICS	LAND USE POLICY
EUROPEAN REVIEW OF AGRICULTURAL ECONOMICS	SCIENCE OF THE TOTAL ENVIRONMENT
RESOURCE AND ENERGY ECONOMICS	RENEWABLE SUSTAINABLE ENERGY REVIEWS
JOURNAL OF FOREST ECONOMICS	
ENVIRONMENTAL AND RESOURCE ECONOMIC REVIEW	
ECOLOGICAL ECONOMICS	
JOURNAL OF ENVIRONMENTAL ECONOMICS AND MANAGEMENT	
AGRICULTURAL ECONOMICS	
TOURISM ECONOMICS	
FOREST POLICY AND ECONOMICS	

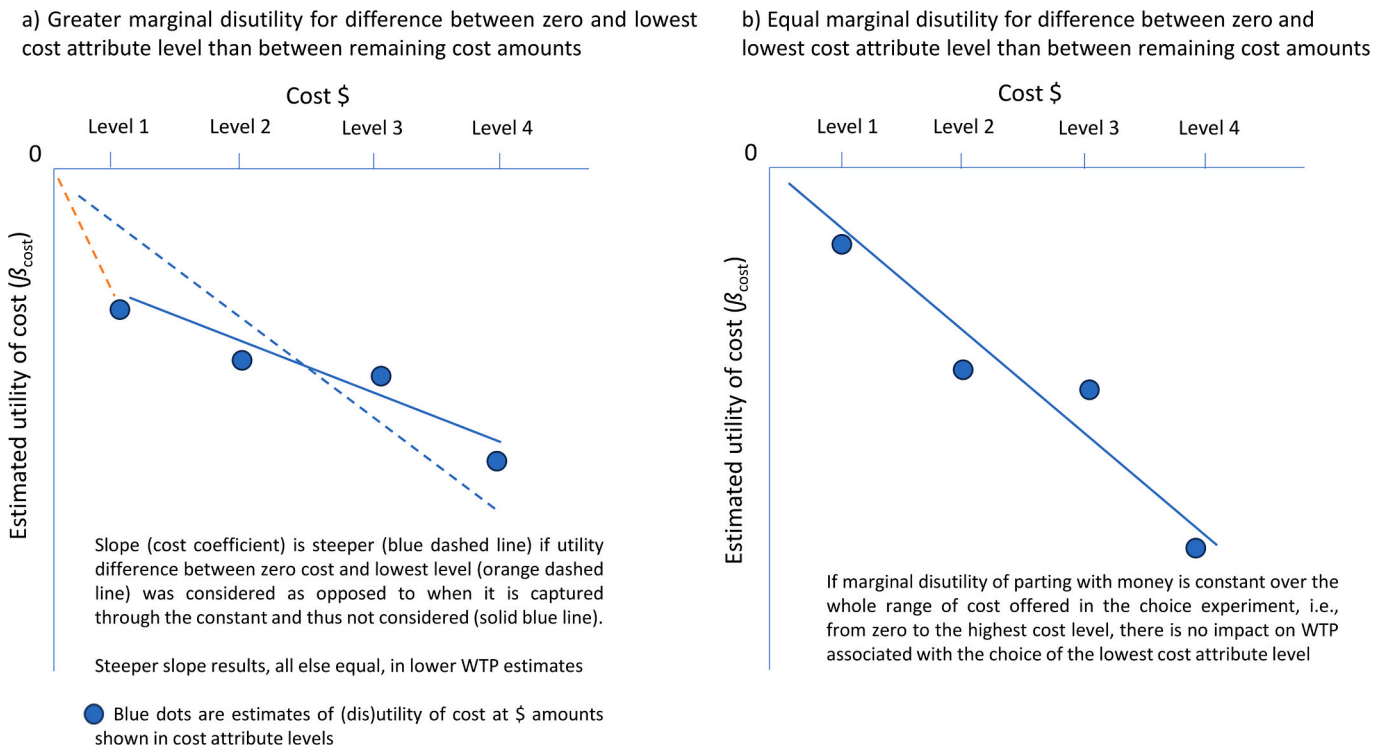


Fig. A1. Illustration of potential impact of choice of lowest cost attribute level on willingness to pay (WTP).
Note: Blue dots represent estimates of (disutility) of the cost coefficient at $ amounts shown in cost attribute levels.

The following summarises the main steps to take to identify whether willingness to pay (WTP) estimates overshoot or not in a study, based on the criterion proposed in this paper (see Section 3.2). The steps are applied to two example research papers included in our sample of studies:

- Glenk, K. and Colombo, S. (2011). How sure can you be? A Framework for considering delivery uncertainty in benefit assessment based on stated preference methods. *Journal of Agricultural Economics*, 62(1): 25–46
- Colombo, S., Calatrava-Requena, J., Hanley, N. (2007). Testing choice experiment for benefit transfer with preference heterogeneity. *American Journal of Agricultural Economics*, 89(1): 135–151

Overview of the steps:

Step 1: for each non-monetary attribute, identify the attribute level that is likely to yield the greatest utility. Note that this is not necessarily the highest attribute level. To identify the attribute level yielding the greatest utility, it may be required to look at the model results in order to observe signs of coefficients and understand relative preferences for attribute levels. For example, an attribute "Number of species lost" may yield a negative coefficient; in this case, the lowest attribute level would be the one yielding the greatest utility. Record the attribute level and the corresponding marginal change. This would be a number for continuous attributes, or a value of 1 for a dummy variable.

Step 2: for each non-monetary attribute, identify the corresponding marginal WTP estimate for the relevant marginal change.

Step 3: using the information collected in steps 1 and 2, calculate the WTP for each attribute level yielding the greatest utility and sum to derive WTP for the bundle of attribute levels yielding the greatest utility.

Step 4: determine the threshold value that defines if a study exhibits overshooting of WTP or not. To do that, compute the difference between the maximum cost attribute level and the minimum cost attribute level (or zero in case that a study has also negative cost levels). See Section 3.2 for a detailed explanation.

Step 5: compare the WTP value computed in Step 3 with the threshold identified in Step 4. If the former exceeds the latter, conclude that the study's WTP estimates overshoot (based on the criterion proposed in this paper).

Application of the steps to Glenk and Colombo (2011):

Step 1:

Attributes	Attribute levels	Attribute level yielding greatest utility
Attribute 1: Annual reduction in net emissions from Scotland (%)	2,4,6,8	8
Attribute 2: Dummy variable taking 1 for 'slight decrease in on-farm employment' as a proxy of impacts on rural viability and 0 for 'no change'	0,1	0
Attribute 3: Dummy variable taking 1 for 'improvement of farmland bird habitat' as a proxy for impacts on biodiversity and 0 for 'no change'	0,1	1

Step 2:

Attributes	Mean marginal WTP for relevant marginal changes Based on Table 3 in Glenk and Colombo (2011), Version 1
Attribute 1	£12.1
Attribute 2	-£20.4
Attribute 3	£30.9

Step 3:

WTP for the bundle yielding highest utility = (£12.1*8) + (- £20.4*0) + (£30.9*1) = £96.8 + 0 + £30.9 = £127.7

Step 4:

Cost attribute levels considered in the study: £ 5, 10, 25, 50, 100, 200

Highest cost attribute level: £200

Lowest cost attribute level (not considering fixed cost of £0 in status quo option): £5

Difference between highest and lowest level: £200 - £5 = £195

Step 5:

Compare the results of Step 3 (£127.7) and Step 4 (£195). Given that the former is lower than the latter, the conclusion (based on the criterion proposed) is that WTP estimates **DO NOT OVERSHOOT**. Note: this does not mean WTP estimates are unbiased.

Application of the steps to Colombo et al. (2007):

Step 1:

Genil area sample:

Attributes	Attribute levels	Attribute level yielding greatest utility
Attribute 1: Landscape change: desertification of the semiarid areas (dummy variables for non-status quo levels)	Degradation (status quo level), Slight improvement, Big Improvement	1 (Big improvement)
Attribute 2: Surface and ground water quality (dummy variable for non-status quo)	Low (status quo level), Medium, High	1 (High)
Attribute 3: Flora and Fauna quality	Poor (status quo level), Medium, Good	1 (Good)
Attribute 4: Agricultural jobs created (number)	0 (status quo level), 100, 200	200
Attribute 5: Area of project execution (km^2)	0 (status quo level), 330, 660, 990	990

Guadajoz area sample:

Attributes	Attribute levels	Attribute level yielding greatest utility
Attribute 1: Landscape change: desertification of the semiarid areas (dummy variables for non-status quo levels)	Degradation (status quo level), Slight improvement, Big Improvement	1 (Big improvement)
Attribute 2: Surface and ground water quality (dummy variable for non-status quo)	Low (status quo level), Medium, High	1 (High)
Attribute 3: Flora and Fauna quality	Poor (status quo level), Medium, Good	1 (Good)
Attribute 4: Agricultural jobs created (number)	0 (status quo level), 65, 130	130
Attribute 5: Area of project execution (km^2)	0 (status quo level), 154, 308, 462	462

Step 2:

Genil area:

Attributes	Mean marginal WTP for relevant marginal changes
Attribute 1	€ 23.50
Attribute 2	€ 21.82
Attribute 3	€ 13.81
Attribute 4	€ 0.104
Attribute 5	€ 0 (not significant)

Guadajoz area:

Attributes	Mean marginal WTP for relevant marginal changes
Attribute 1	€ 23.76
Attribute 2	€ 26.30
Attribute 3	€ 13.75
Attribute 4	€ 0.161
Attribute 5	€ 0.031

Step 3:

WTP for the bundle yielding greatest utility (Genil):

= (€23.50*1) + (€21.82*1) + (€13.81*1) + (€0.104*200) + (€0*990) = €79.93

WTP for the bundle yielding greatest utility (Guadajoz):

= (€23.76*1) + (€26.30*1) + (€13.75*1) + (€0.161*130) + (€0.031*462) = €99.06

Step 4:

Both Genil and Guadajoz areas:

Cost attribute levels considered in the study: € 6.01, 12.02, 18.03, 24.04, 30.05, 36.06

Highest cost attribute level: €36.06

Lowest (non-zero) cost level: €6.01

Difference between highest and lowest level €36.06 - €6.01 = €30.05

Step 5:

Genil area:

Compare the results of Step 3 (€79.93) and Step 4 (€30.05). Given that the former is greater than the latter, the conclusion (based on the criterion proposed) is that WTP estimates **OVERSHOOT.**

Guadajoz area:

Compare the results of Step 3 (€99.06) and Step 4 (€30.05). Given that the former is greater than the latter, the conclusion (based on the criterion proposed) is that WTP estimates **OVERSHOOT.**

Box A1. Step-by-step guide to applying the face validity check for two example publications (Glenk and Colombo, 2011).

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecolecon.2024.108160.

References

Ahtiainen, H., Pouta, E., Zawadzki, W., Tienhaara, A., 2023. Cost vector effects in discrete choice experiments with positive status quo cost. J. Choice Model. 47, 100401.

Alemu, M., Mørkbak, M., Olsen, S., Jensen, C., 2013. Attending to the reasons for attribute non-attendance in choice experiments. Environ. Resour. Econ. 54, 333–359.

Andreoni, J., 1990. Impure altruism and donations to public goods: a theory of warmglow giving. Econ. J. 100, 464–477.

Bateman, I.J., Carson, R.T., Day, B., Hanemann, M., Hanley, N., Hett, T., Jones-Lee, M., Loomes, G., Mourato, S., Özdemiroglu, E., Pearce, D.W., 2002. Economic Valuation with Stated Preference Techniques: A Manual. Edward Elgar, Cheltenham, USA.

Bishop, R.C., Boyle, K.J., 2019. Reliability and validity in nonmarket valuation. Environ. Resour. Econ. 72, 559–582.

Brouwer, R., Dekker, T., Rolfe, J., Windle, J., 2010. Choice certainty and consistency in repeated choice experiments. Environ. Resour. Econ. 46, 93–109.

Brown, T.C., Champ, P.A., Bishop, R.C., McCollum, D.W., 1996. Which response format reveals the truth about donations to a public good. Land Econ. 72 (2), 152–166.

Cameron, A.C., Trivedi, P.K., 2010. Microeconometrics Using Stata. Stata Press, Texas.

Carlsson, F., Frykblom, P., Lagerkvist, C.J., 2005. Using cheap-talk as a test of validity in choice experiments. Econ. Lett. 89 (2), 147–152.

Carlsson, F., Kataria, M., Krupnick, A., Lampi, E., Lofgren, A., Qin, P., Sterner, T., 2013. The truth, the whole truth, and nothing but the truth: a multiple country test of an oath script. J. Econ. Behav. Organ. 89, 105–121.

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

Caussade, S., Ortúzar, J., Rizzi, L.I., Hensher, D.A., 2005. Assessing the influence of design dimensions on stated choice experiment estimates. Transp. Res. B 39, 621–640.

Chien, Y., Huang, C.J., Shaw, D., 2005. A general model of starting point bias in double-bounded dichotomous contingent valuation surveys. J. Environ. Econ. Manag. 50 (2), 363–377.

Colombo, S., Glenk, K., Rocamora-Montiel, B., 2016. Analysis of choice inconsistencies in on-line choice experiments: impact on welfare measures. Eur. Rev. Agric. Econ. 43 (2), 271–302.

Colombo, S., Budziński, W., Czajkowski, M., Glenk, K., 2022. The relative performance of *ex-ante* and *ex-post* measures to mitigate hypothetical and strategic bias in a stated preference study. J. Agric. Econ. 73 (3), 845–873.

Comerford, D.A., Hanley, N., 2017. The External Validity of Consequential Stated Preference Studies: A Comment. Paper 2017-02. University of St. Andrews Discussion papers in Environmental Economics.

Crastes Dit Sourd, R., 2023. A new empirical approach for mitigating exploding implicit prices in mixed logit models. Am. J. Agric. Econ. https://doi.org/10.1111/ajae.12367.

Cummings, R.G., Taylor, L.O., 1999. Unbiased value estimates for environmental goods: a cheap talk design for the contingent valuation method. Am. Econ. Rev. 89 (3), 649–665.

Czajkowski, M., Vossler, C.A., Budziński, W., Wiśniewska, A., Zawojska, E., 2017. Addressing empirical challenges related to the incentive compatibility of stated preference methods. J. Econ. Behav. Organ. 142, 47–63.

de Magistris, T., Azucena, G., Nayga, R.M., 2013. On the use of honesty priming tasks to mitigate hypothetical bias in choice experiments. Am. J. Agric. Econ. 95 (5), 1136–1154.

DeShazo, J.R., Fermo, G., 2002. Designing choice sets for stated preference methods: the effects of complexity on choice consistency. J. Environ. Econ. Manag. 44, 123–143.

Entem, A., Lloyd-Smith, P., Adamowicz, W.V.L., Boxall, P.C., 2021. Using inferred valuation to quantify survey and social desirability bias in stated preference research. Am. J. Agric. Econ. 104 (4), 1224–1242.

Faccioli, M., Glenk, K., 2022. More in good condition or less in bad condition? Valence-based framing effects in environmental valuation. Land Econ. 98 (2), 314–336.

Glenk, K., Colombo, S., 2011. Designing policies to mitigate the agricultural contribution to climate change: an assessment of soil based carbon sequestration and its ancillary effects. Climatic Change 105, 43–66. https://doi.org/10.1007/s10584-010-9885-7.

Glenk, K., Martin-Ortega, J., Pulido-Velazquez, M., Potts, J., 2015. Inferring attribute non-attendance from discrete choice experiments: implications for benefit transfer. Environ. Resour. Econ. 60, 497–520.

Glenk, K., Meyerhoff, J., Akaichi, F., Martin-Ortega, J., 2019. Revisiting cost vector effects in discrete choice experiments. Resour. Energy Econ. 57, 135–155.

Haltia, E., Kuuluvainen, J., Ovaskainen, V., Pouta, E., Rekola, M., 2009. Logit model assumptions and estimated willingness to pay for forest conservation in southern Finland. Empir. Econ. 37, 681–691.

Hanley, N., Czajkowski, M., 2019. The role of stated preference valuation methods in understanding choices and informing policy. Rev. Environ. Econ. Policy 13 (2), 248–266.

Hess, S., Beharry-Borg, N., 2012. Accounting for latent attitudes in willingness-to-pay studies: the. 3. Case of coastal water quality improvements in Tobago. Environ. Resour. Econ. 52 (1), 109–131.

Hess, S., Stathopoulos, A., Campbell, D., O'Neill, V., Caussade, S., 2013. It's not that I don't care, I just don't care very much: confounding between attribute non-attendance and taste heterogeneity. Transportation 40 (3), 583–607.

Howard, G., Roe, B.E., Nisbet, E.C., Martin, J.F., 2017. Hypothetical Bias mitigation techniques *in* choice experiments*:* do cheap talk and honesty priming effects fade *with* repeated choices*?* J. Assoc. Environ. Resour. Econ. 4 (2), 543–573.

Johnston, R.J., Boyle, K.J., Adamowicz, W., Bennett, J., Brouwer, R., Cameron, T.A., Hanemann, W.M., Hanley, N., Scarpa, R., Tourangeau, R., Vossler, C.A., 2017. Contemporary guidance for stated preference studies. J. Assoc. Environ. Resour. Econ. 4 (2), 319–405.

Kanninen, B., 1995. Bias in discrete response contingent valuation. J. Environ. Econ. Manag. 28, 114–125.

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., 2014. Augmenting short cheap talk scripts with a repeated opt-out reminder in choice experiment surveys. Resour. Energy Econ. 37, 39–63.

Liebe, U., Glenk, K., Oehlmann, M., Meyerhoff, J., 2015. Does the use of mobile devices (tablets and smartphones) affect survey quality and choice behaviour in web surveys? J. Choice Model. 14, 17–31.

Lindhjem, H., Navrud, S., 2011. Are internet surveys an alternative to face-to-face interviews in contingent valuation? Ecol. Econ. 70 (9), 1628–1637.

Lopez-Becerra, E.I., Alcon, F., 2021. Social desirability bias in the environmental economic valuation: an inferred valuation approach. Ecol. Econ. 184, 106988 https://doi.org/10.1016/j.ecolecon.2021.106988.

Mariel, P., Hoyos, D., Meyerhoff, J., Czajkowski, M., Dekker, T., Glenk, K., Jacobsen, J.B., Liebe, U., Olsen, S.B., Sagebiel, J., Thiene, M., 2021. Environmental Valuation with Discrete Choice Experiments: Guidance on Design, Implementation and Data Analysis. Springer Nature.

Martin, S., 2019. The Kaldor-Hicks potential compensation principle and the constant marginal utility of income. Rev. Ind. Organ. 55, 493–513.

Meyerhoff, J., Oehlmann, M., 2023. The performance of full versus partial profile choice set designs in environmental valuation. Ecol. Econ. 204, 107665.

Meyerhoff, J., Oehlmann, M., Weller, P., 2015. The influence of design dimensions on stated choices in an environmental context. Environ. Resour. Econ. 61, 385–407.

Meyerhoff, J., Mariel, P., Bertram, C., Rehdanz, K., 2017. Matching preferences or changing them? The influence of the number of choice alternatives. In: 23rd Annual Conference of the European Association of Environmental and Resource Economists. Athens, Greece.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., The PRISMA Group, 2009. Preferred reporting items for systematic reviews and Meta-analyses: the PRISMA statement. PLoS Med. 6 (7), e1000097.

Mørkbak, R.M., Christensen, T., Gyrd-Hansen, D., 2010. Choke Price Bias in choice experiments. Environ. Resour. Econ. 45 (4), 537–551.

OECD, 2017. OECD Guidelines on Measuring Trust. OECD Publishing, Paris. https://doi.org/10.1787/9789264278219-en.

Parsons, G.R., Myers, K., 2016. Fat tails and truncated bids in contingent valuation: an application to an endangered shorebird species. Ecol. Econ. 129, 210–219.

Penn, J.M., Hu, W., 2018. Understanding hypothetical bias: an enhanced meta-analysis. Am. J. Agric. Econ. 100 (4), 1186–1206.

Petrolia, D., Interis, M., Hwang, J., 2014. America's wetland? A national survey of willingness-to-pay for restoration of Louisiana's coastal wetlands. Mar. Resour. Econ. 29 (1), 17–37.

Rakotonarivo, O.S., Schaafsma, M., Hockley, N., 2016. A systematic review of the reliability and validity of discrete choice experiments in valuing non-market environmental goods. J. Environ. Manag. 183, 98–109.

Rao, A.R., Monroe, K.B., 1989. The effect of price. Brand name. And store name on buyers' perceptions of product quality: an integrative review. J. Mark. Res. 36, 351–357.

Ready, R.C., Hu, D., 1995. Statistical approaches to the fat tail problem for dichotomous choice contingent valuation. Land Econ. 71 (4), 491–499.

Riera, P., Giergiczny, M., Penuelas, J., Mahieu, P.-A., 2012. A choice modelling case study on climate change involving two-way interactions. J. For. Econ. 18, 345–354.

Rollins, C., 2023. Investigating cost non-attendance as a driver of inflated welfare estimates in mixed-logit models. J. Agric. Econ. https://doi.org/10.1111/1477-9552.12558.

Scarpa, R., Campbell, D., Hutchinson, W.G., 2007. Benefit estimates for landscape improvements: sequential Bayesian design and respondents' rationality in a choice experiment study. Land Econ. 83 (4), 617–634.

Scarpa, R., Gilbride, T., Campell, D., Hensher, D.A., 2009. Modelling attribute non-attendance in choice experiments for rural landscape valuation. Eur. Rev. Agric. Econ. 36 (2), 151–174.

Schaafsma, M., Brouwer, R., 2020. Substitution effects in spatial discrete choice experiments. Environ. Resour. Econ. 75, 323–349.

Tonsor, G.T., Shupp, R.S., 2011. Cheap talk scripts and online choice experiments: looking beyond the mean. Am. J. Agric. Econ. 93 (4), 1015–1031.

Train, K., Weeks, M., 2005. Discrete choice models in preference space and willingness-to-pay space. In: Scarpa, R., Alberini, A. (Eds.), Application of Simulation Methods in Environmental and Resource Economics. Springer, Dordrecht, pp. 1–16.

Villanueva, A.J., Glenk, K., Rodríguez-Entrena, M., 2017. Protest responses and willingness to accept: Ecosystem Services Providers' preferences towards incentive-based schemes. J. Agric. Econ. 68, 801–821.
Vossler, C.A., Doyon, M., Rondeau, D., 2012. Truth in consequentiality: theory and field evidence on discrete choice experiments. Am. Econom. J. Microeconom. 4, 145–171.
Welling, M., Zawojska, E., Sagebiel, J., 2022. Information, consequentiality and credibility in stated preference surveys: a choice experiment on climate adaptation. Environ. Resour. Econ. 82, 257–283.
Weng, W., Morrison, M.D., Boyle, K.J., Boxall, P., Rose, J., 2020. Effects of the number of alternatives in public good discrete choice experiments. Ecol. Econ. 182, 106904.
Zawojska, E., Bartczak, A., Czajkowski, M., 2019. Disentangling the effects of policy and payment consequentiality and risk attitudes on stated preferences. J. Environ. Econ. Manag. 93, 63–84.
Zhang, J., Adamowicz, W.L., 2011. Unraveling the choice format effect: a context-dependent random utility model. Land Econ. 87, 730–743.