The relative performance of ex-ante and ex-post measures to mitigate hypothetical and strategic bias in a stated preference study

Colombo, Sergio; Budziski, Wiktor; Czajkowski, Mikoaj; Glenk, K

Published in:
Journal of Agricultural Economics

DOI:
10.1111/1477-9552.12484

First published: 29/03/2022

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

Received: 17 March 2021 | Revised: 22 February 2022 | Accepted: 8 March 2022

DOI: 10.1111/1477-9552.12484

ORIGINAL ARTICLE

Journal of Agricultural Economics

The relative performance of *ex-ante* and *ex-post* measures to mitigate hypothetical and strategic bias in a stated preference study

Sergio Colombo[1,2] | **Wiktor Budziński**[3] | **Mikołaj Czajkowski**[3] | **Klaus Glenk**[4]

[1]Department of Agricultural Economics, IFAPA, Granada, Spain

[2]WEARE-Water, Environmental and Agricultural Resources Economics Research Group, University of Cordoba, Cordoba, Spain

[3]Department of Economics, University of Warsaw, Warsaw, Poland

[4]Scotland's Rural College, SRUC, Edinburgh, Spain

Correspondence
Sergio Colombo, Department of Agricultural Economics, IFAPA, Granada, Spain.
Email: sergio.colombo@juntadeandalucia.es

Funding information
Andalusian Institute of Agricultural Economics and Training (IFAPA) and co-funded by FEDER fund, Grant/Award Number: PP.AVA.AVA2019.031; National Science Centre of Poland, Grant/Award Number: 2017/25/B/HS4/01076; National Science Centre of Poland, Grant/Award Number: 2016/21/N/HS4/02094

Abstract

Bias related to the hypothetical setting remains controversial regarding the reliability and validity of value estimates from discrete choice experiments (DCEs). This has motivated a large body of literature to investigate approaches for mitigating hypothetical and strategic bias. Our study provides further evidence to inform this debate by testing whether a combination of *ex-ante* or *ex-post* mitigation strategies might be effective in reducing bias in DCEs. Specifically, we employ individual and multiple *ex-ante* reminders alongside an *ex-post* data treatment and analyse how their individual or joint use affects willingness to pay (WTP) estimates. The econometric analysis makes use of innovative semi-parametric logit-mixed logit in addition to the state-of-the-art mixed logit model. The empirical case study focuses on preferences for the environmental and social impacts of organic olive production. By comparing the three experimental treatments with a control treatment, we test whether *ex-ante* cheap talk, a reminder of the project's relative spatial extent, or a combination of both affect stated WTP. In addition, we use an *ex-post* data treatment to correct WTP estimates. WTP estimates of treatments related to *ex-ante* mitigation strategies did not differ significantly from those obtained from a control treatment with standard budget constraint reminders. However, the *ex-post* approach results in a significant reduction in mean WTP estimates and is used to investigate whether the observed choice inconsistencies are due to unintentional errors or strategic behaviour. We argue that *ex-post* mechanisms deserve greater attention and highlight the need to distinguish strategic behaviour from other sources of hypothetical bias.

KEYWORDS

choice experiment, hypothetical bias, mitigation strategies, organic olive grove cultivation, Southern Spain, strategic bias

JEL CLASSIFICATION

C12; C93; D12; Q01; Q20; Q51

1 | INTRODUCTION

Over the past decade, stated preferences (SP) have become the main valuation approach used to identify the value of environmental goods (Hanley & Czajkowski, 2019). However, SP studies have been widely criticised for failing validity and reliability tests (Bishop & Boyle, 2019). As an increasingly popular SP method, the potential for discrete choice experiments (DCEs) to provide credible information for decision-making must still be scrutinised and procedures should be explored to improve the validity of welfare estimates derived from DCEs (Johnston et al., 2017).

The academic community has paid much attention to the issue of hypothetical bias (HB), given the inherent hypothetical nature of the valuation task, which has been frequently observed to distort responses in DCEs compared with behaviour in real markets (Haghani et al., 2021; Loomis, 2014; Lusk & Schroeder, 2004; Vossler et al., 2012). HB in DCEs may arise for various reasons, such as respondents not caring about the outcome of the survey (e.g., if they think the survey will not matter for any policy decision, regarding it as inconsequential), or, conversely, believing that by misrepresenting their true preferences they can influence the outcome of the survey to their advantage (strategic bias, SB). This has given rise to the research considering incentive compatibility (Carson and Groves, 2007) and consequentiality (Vossler et al., 2012) of SP methods.

Incentive compatible surveys are surveys where respondents' dominant strategy is to reveal their preferences truthfully (Carson and Groves, 2007). Achieving incentive compatibility in DCEs is difficult in practice because it requires choice set independence and one-to-one matching between the project identified in a choice set and the policies that might be implemented (Vossler et al., 2012). Theoretically, this may be achieved in a single binary choice question format (Gibbard, 1973; Satterthwaite, 1975) but not in the multinomial choice sequence typically employed in DCE. However, as Zawojska and Czajkowski (2017) pointed out, a relatively small body of empirical literature addresses the problem of bias resulting from using more than two alternatives or more than one choice task per respondent. If the resulting bias is relatively small, the statistical efficiency of more elaborate elicitation formats could outweigh the bias resulting from being incentive incompatible.

The elicitation of true preferences also requires that the survey is perceived by respondents as consequential (Carson and Groves, 2007). As demonstrated by Zawojska et al.

(2019), survey consequentiality is met when respondents view their responses as influencing the policy decision and when they care about the outcomes (policy consequentiality). Additionally, respondents must perceive the payment associated with the policy as consequential—they must believe that in case a policy is implemented, they will have to pay the amount specified (payment consequentiality). Vossler et al. (2012) and Groothuis et al. (2017) found that consequentiality is established when there is a positive probability of the survey being consequential, and Needham and Hanley (2020) observed a monotonic relationship between the degree of perceived consequentiality and WTP. Perceived consequentiality differs also according to prior knowledge of the good (Needham and Hanley, 2020), elicitation format (Lloyd-Smith et al., 2019) and can be affected by endogeneity issues (Groothuis et al., 2017). Also, perceived consequentiality is often not easy to measure (Czajkowski et al., 2017; Needham and Hanely, 2020).

HB and SB are both clearly linked to incentive compatibility and survey consequentiality; however, their relationships requires further discussion.[1] Although fully consequential surveys preclude HB, and fully hypothetical surveys preclude SB, we argue that most of the SP valuation studies conducted in the field are not perceived by respondents as fully hypothetical or fully consequential, but rather as something in between. It is sufficient that respondents expect that either the payment or provision of a good is not fully consequential to introduce the possibility of HB (Zawojska et al., 2019). However, if the survey design is perceived as consequential, it has the possibility of SB—respondents may want to misrepresent their preferences to influence the policy decision or payment. As a result, HB and SB are intertwined and often co-exist in the same SP setting.

It is widely accepted that HB may lead respondents to overstate their willingness to pay (WTP). This is particularly likely in the context of public goods where respondents might be keen to demonstrate support for the enhanced public good provision or might want to represent themselves as 'doing the right or good thing', especially if they do not expect that they would have to personally bear the specified cost (Johansson-Stenman and Svedsäter, 2012; Morrison and Brown, 2009). Therefore, a large body of DCE literature has focused on mechanisms to reduce HB or on explaining its underlying behavioural reasons (Grebitus et al., 2013; Jacquemet et al., 2013; Kang et al., 2011) when investigated in relation to private goods (Doyon et al., 2015; Liebe et al., 2019; Lusk and Schroeder, 2004; Moser et al., 2014; Tonsor and Shupp, 2011) and public goods (Carson and Groves, 2007; Czajkowski et al., 2017; Zawojska et al., 2019). Despite these efforts, mitigating HB remains a concern for SP valuation studies (Murphy, Allen, et al., 2005; Loomis, 2014). In a recent review, Haghani et al. (2021) concluded that in health-related choice experiments negligible degrees of HB are observed; experiments in consumer behaviour and transport domains show significant degrees of HB, and environmental valuation studies provide mixed evidence about the presence of this bias. Overall, the results of the studies concerned with HB and SB show that caution should be used when informing policies about public good provision using information based on DCEs.

Approaches to reduce HB in SP-based valuations can be classified into *ex-ante* and *ex-post* mitigation strategies.[2] *Ex-ante* approaches aim to reduce the bias at the survey design stage by emphasising the consequences of respondents' choices, for example, in terms of additional

[1]We are grateful to an anonymous referee for pointing this out and for their contribution in improving previous versions of our paper.

[2]Note that henceforth we are referring to reducing HB unless we specifically discuss SB. However, we acknowledge that it is possible that *ex-ante* and *ex-post* mitigation strategies also addressed SB (to some extent). We have decided to follow this assumption, because this was the term used in the studies that employed *ex-ante* and *ex-post* mitigation measures, and because the primary ways of correcting for SB concern the design of a study (e.g., the number of alternatives and choice tasks). For a comprehensive review of *ex-ante* and *ex-post* mitigation strategies, see Loomis (2011, 2014).

payments, or by reminding them to behave as they would in a real choice or purchasing situation (e.g., to consider budget restrictions, the existence of substitutes, or to avoid socially desirable responses). Amongst the *ex-ante* tools often employed, Cheap Talk Scripts (CTSs) have been widely used to ask respondents to consider responding as if payments were real. Despite the simplicity of CTSs, empirical evidence regarding their effectiveness is mixed. Some studies found CTSs to be successful in mitigating HB (Carlsson et al., 2005; Chowdhury et al., 2011; Ladenburg and Olsen, 2014; List et al., 2006; Tonsor and Shupp, 2011), while others observed no effects (Blumenschein et al., 2008; Bosworth and Taylor, 2012; Moser et al., 2014; Varela et al., 2014). Multiple reasons underpin these results. Generally, the effectiveness of CTSs tends to increase as the amount of CTSs increases (Brown et al., 2003; Murphy, Stevens, et al., 2005). The performance of CTSs also varies depending on respondents' previous experience or knowledge (Champ et al., 2009; Lusk, 2003), and they work better with less experienced individuals. Several studies found CTSs to be effective only for specific groups of people (Aadland and Caplan, 2003; Ami et al., 2011; Barrage and Lee, 2010; Champ et al., 2009) or for public goods (Penn and Hu, 2019).

Ex-post approaches address HB at the data analysis stage by means of procedures that screen the data for implausible responses, often based on responses to questions asked after the valuation tasks. The oldest *ex-post* approach is the follow-up certainty question, which asks respondents to state how certain they are about their choices (Akter and Bennett, 2013; Blumenschein et al., 1998; Blomquist et al., 2009; Champ et al., 2009; Johannesson et al., 1998; Ready et al., 2010). Another *ex-post* approach asks respondents to state their maximum WTP for the good in question in order to detect their preference uncertainty (Bush et al., 2009) or to identify choice inconsistencies (Colombo et al., 2016). An *ex-post* approach that does not rely on follow-up questions relies on the combination of data from revealed preference studies (if possible) with SP data (Azevedo et al., 2003; Brooks and Lusk, 2010) or the use of revealed preference data to calibrate stated WTP (Fox et al., 1998). *Ex-post* approaches generally find that follow-up questions can be used to obtain WTP estimates that better reflect true preferences. However, as several authors have pointed out, a drawback of *ex-post* approaches with follow-up questions is that responses need to be adequately calibrated, given that it is necessary to determine at which level hypothetical decisions correspond to a real decision. Incorrect response calibration may result in further bias (Beck et al., 2016).

Simultaneous application of more than one mitigation technique may enhance HB reduction. For example, Ladenburg and Olsen (2014) and Varela et al. (2014) combined two *ex-ante* approaches by augmenting CTS with an opt-out reminder. However, this was only found to be effective in the case of Ladenburg and Olsen (2014). As Loomis (2014) pointed out, *ex-ante* and *ex-post* approaches may also be combined in a single study. A few studies have examined both approaches simultaneously in a Contingent Valuation (CV) setting (Blumenschein et al., 2008; Champ et al., 2009; Morrison and Brown, 2009; Whitehead and Cherry, 2007). The results of these studies suggest that *ex-post* approaches might be better positioned for reducing HB than *ex-ante* approaches. Whitehead and Cherry (2007) concluded that WTP estimates are similar when either *ex-ante* or *ex-post* approaches are employed, suggesting that the approaches should be considered as complements rather than substitutes. However, Blumenschein et al. (2008), Champ et al. (2009), and Morrison and Brown (2009) found that *ex-post* mitigation measures were more effective in reducing HB. An exception is Broadbent (2014), who investigated the joint effect of using both *ex-ante* and *ex-post* approaches in a DCE setting. In his study, concerning a quasi-public good, he found that marginal WTP between hypothetical and real choices did not differ statistically and concluded that in this case, it is unnecessary to employ any mitigation instruments. To our knowledge, the effect of using both approaches in the case of a DCE study on public goods has not yet been investigated.

We address this research gap by drawing on data from a DCE on the environmental and social impacts of organic olive oil production to investigate the joint effect of *ex-ante* and *ex-post* approaches on mitigating HB. For the *ex-ante* approach, we test the sensitivity of the

WTP values to different CTS formats. In addition to the typical CTS that informs respondents about a common propensity to exaggerate stated WTP, we also consider a CTS that refers to the proposed project's relative extent (scale) and test whether either of the CTS formats affects the extent of HB. Additionally, we test whether the use of a combination of both CTSs further affects the stated WTP. Regarding the *ex-post* approach, we follow the methodology proposed by Colombo et al. (2016), who reduced HB by allowing respondents to revise those choices in the DCE that were found to be inconsistent with responses to a follow-up question. Owing to multiple experimental treatments, our empirical data have four times the number of observations compared to Colombo et al. (2016). Importantly, the *ex-post* mitigation strategy does not require the application of any calibration function (or concomitant assumptions). Furthermore, the *ex-post* approach employed offers an opportunity to classify choices affected by either HB for SB, or by any other reason for respondents not responding truthfully. The *ex-post* approach therefore provides us with a tool to disentangle the effect of these biases. Finally, to shed light on whether the mixed results observed in the literature on the performance of various HB mitigation strategies are due to the modelling approach, we use both a standard multinomial mixed (random parameters) logit model (MXL; Revelt and Train, 1998) and the more recent semi-parametric logit-mixed logit (LML; Train, 2016) model. The LML is arguably a more flexible approach that allows for the estimation of the shape of the preference heterogeneity distribution without imposing restrictive assumptions regarding its parametric specification. This is one of the first applications of the LML model, thus providing further insights regarding its potential superiority over the standard MXL model (Bansal et al., 2018a; Bazzani et al., 2018; Franceschinis et al., 2017).

The remainder of this paper is organised as follows. Section 2 introduces the study design, focusing on the *ex-ante* and *ex-post* procedures employed to mitigate HB, and provides an overview of the case study. In Section 3, we outline the methodology used for the data analysis. We then present the results and discuss the implications for decision making in Section 4, ending the article with a set of conclusions presented in Section 5.

2 | STUDY DESIGN

An online questionnaire was developed to elicit respondents' preferences regarding the environmental and social impacts of olive growing in the sloping areas of the Andalucía region in southern Spain. In the survey, respondents were told about olives using short and clear pieces of information to engage and keep respondents' attention. We employed several graphical illustrations to describe the olive growing production systems and their environmental and social impacts, which constitute the attributes of the DCE. Appendix A, online, provides the text version of the survey. We clearly explained that the four different olive growing production systems (marginal, traditional, intensive and super-intensive) are associated with specific environmental and social impacts, and that among the four systems, this study focuses on marginal olive production, which has a high potential to enhance public good provision. The DCE attributes were: (i) reducing the impact of climate change, (ii) enhancing biodiversity, (iii) reducing the risk of the pollution of water resources, (iv) limiting soil erosion, (v) increasing agricultural employment, and (vi) increasing tax. All attributes, except the tax increase, were treated as qualitative and coded as dummy variables. In all cases, detailed information about the level each attribute takes at the time of the survey and could take in the future through project implementation was provided to respondents with graphical illustrations and a short text. To help comprehension and consideration of the attributes and levels, we also provided this information in a simple table, which was available on demand throughout completion of the choice tasks.

Table 1 lists the attributes and their respective levels. An example choice card is shown in Figure 1. Six choice cards were presented to each respondent. Each choice task had three

TABLE 1 Attributes and levels of the choice experiment

Attribute	Levels
Tackling climate change	**Low**
	Medium
	High
Biodiversity	**Low**
	Medium
	High
Risk of pollution of water resources	**High**
	Moderate
	Low
Soil erosion	**High**
	Moderate
	Low
Agricultural employment	**0%**, +5% +10%
Tax	**0**, 2, 7, 14, 23, 35, 51 €/year

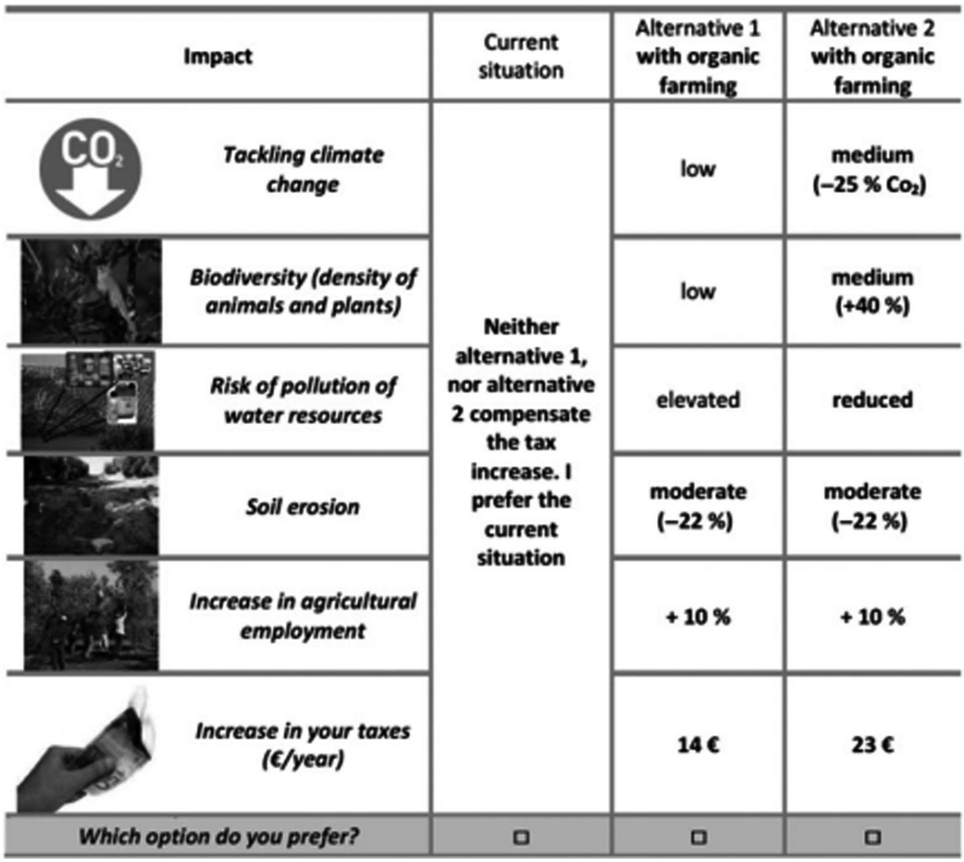
Note: Levels in the current situation are shown in bold.

alternatives. Two alternatives described the impacts of potential changes in olive growing in the region and were associated with a tax increase. A third alternative represented the status quo and was available at no extra cost. A fractional factorial experimental design was determined by minimising the D-error for an MNL specification using Bayesian techniques with priors determined from an earlier pilot study involving 30 people.

We aim to test the effect of HB mitigation approaches on the premise that HB is likely to be present, given the public goods context of the DCE (Morrison and Brown, 2009) and the lack of incentive compatibility in the experiment (Carson and Groves, 2007; Vossler et al., 2012).[3] Our working hypothesis is that *ex-ante* and *ex-post* approaches will result in a reduction in the estimated WTP. We interpret a reduction in WTP associated with the use of *ex-ante* and *ex-post* approaches as a mitigation of potential HB. For the experimental treatments associated with the different (combinations of) *ex-ante* devices tested and their combination with an *ex-post* approach, our hypothesis is that (mean) WTP based on these treatments is different from (mean) WTP based on a control treatment in which no additional HB reduction is undertaken.

One of the study's limitations is that we did not include a 'real' treatment that may be used as a benchmark for disclosing respondents' true preferences. In this sense, we cannot be sure of the magnitude of true unbiased WTP values. Therefore, we assume that lower WTP estimates are closer to the unbiased value, given that monotonic preferences are to be expected in the experiment. Better environmental and social conditions are expected to be preferred to

[3]As one referee pointed out, the DCE design used in this experiment is not theoretically incentive compatible and, thus, we would not expect choices to provide reliable estimates of WTP, since respondents have an incentive to strategically misrepresent their preferences. This has been documented empirically in laboratory studies (Meginnis et al., 2021; Silz-Carson et al., 2020), and it is of interest to test the degree to which this occurs in the field. As Zawojska and Czajkowski (2017) stated, a small body of empirical literature addresses bias problems resulting from using more than two options or more than one choice task per respondent (see Weng et al., 2021, for an updated review). If the resulting bias is relatively small, the advantages of the enhanced statistical efficiency of elicitation formats that deviate from theoretical incentive compatibility could outweigh the negative consequences in terms of bias.



Impact	Current situation	Alternative 1 with organic farming	Alternative 2 with organic farming
Tackling climate change	Neither alternative 1, nor alternative 2 compensate the tax increase. I prefer the current situation	low	medium (–25 % Co₂)
Biodiversity (density of animals and plants)		low	medium (+40 %)
Risk of pollution of water resources		elevated	reduced
Soil erosion		moderate (–22 %)	moderate (–22 %)
Increase in agricultural employment		+ 10 %	+ 10 %
Increase in your taxes (€/year)		14 €	23 €
Which option do you prefer?	☐	☐	☐

FIGURE 1 Example choice card

worse conditions, all else being equal. Furthermore, although some authors have used 'non-hypothetical' treatments, we argue that it is never possible to mimic real consumers' behaviour. For instance, Chowdhury et al. (2011), and Broadbent (2014) provided respondents with an initial monetary endowment that they had to spend according to the choices made. Alemu and Olsen (2018) gave respondents a lump sum prior to the DCE and informed them that they were welcome to keep the money for their own use. However, evidence suggests that respondents' behaviour may differ depending on whether the wealth is regarded as a 'windfall' or 'earned' (Cherry et al., 2002). To our knowledge, only Moser et al. (2014) and Liebe et al. (2019) carried out a field experiment in which respondents used their own money, and Liebe et al. (2019) implemented the experiment in an online format. However, these two studies were made possible because they analysed preferences for existing consumer (market) goods of relatively low value, and is not practical for non-market goods. Furthermore, in the case of public goods, where a tax is used as the payment vehicle, providing respondents with a monetary incentive would equate to some form of a 'tax-rebate', which is implausible and may create distrust or confusion for respondents.

To appraise *ex-ante* approaches to mitigate HB, four different versions (Treatments 1–4 [T1–4]) of the questionnaire were administered. In all treatments, following Ladenburg and Olsen (2014) and Varela et al. (2014), we carefully described the choice task and included a typical reminder about respondents' budget constraints and the existence of alternative goods they may prefer to consume. Special effort was made to explain the consequentiality of the study in terms of the tax increase associated with the choice of alternatives. As Vossler et al.

(2012) observed, if respondents see the experiment as 'sufficiently' consequential, hypothetical WTP estimates do not differ statistically from real preference WTP estimates.[4] In particular, we informed respondents that the survey was funded by the Andalusian government, which is well known to be the authority with competence for regional agri-environmental policy and which can impose taxes for environmental purposes. In the studied area, olive growing is a deeply subsidised production system (about 95% of the olive grove area receives a subsidy from the EU Common Agricultural Policy), and respondents were aware of this.

We employed the available features of online surveys to encourage respondents to think carefully about their choices. Features included delaying the next steps through the survey to convey important information and providing pop-up windows to explain the choice task and provide definitions of the attribute levels. T1 served as a control treatment in which additional *ex-ante* HB mitigation strategies were absent. In T2, respondents received a CTS that informed them about the likely HB in the choices and explicitly reminded them to consider that the hypothetical nature of the choices may influence their answer and lead to erroneous study conclusions, which may result in the application of a higher tax increase. The CTS script reads:

Please note that previous research shows that respondents sometimes selected an alternative that they would not have chosen if they really had to pay for it. In other words, they chose an alternative without considering the associated costs, because they did not have to pay for them there and then. This could lead us to the wrong conclusions and could even result in the application of a higher tax increase by the Junta de Andalucía. For this reason, we ask you only to choose an alternative situation if you are willing to pay more tax in exchange for the benefits described. Otherwise, simply choose the current situation.

In T3, we included a CTS aimed at considering the project's relative extent. This treatment serves to investigate the likely spatial dimensions of the HB. Given the hypothetical nature of the project described in the DCE, respondents may not focus on the dimension of the proposed changes and may express a WTP measure that considers the entire olive grove area instead of only focusing on the area of marginal olive groves. In this treatment, a pie chart accompanied a textual reminder to convey information about the proportion of marginal olive orchard area affected by the project relative to total agricultural area (8%). The accompanying CTS, which reminded respondents that the tax payments associated with each choice card alternative would only provide the environmental and social benefits described in the affected area, reads as follows:

Please bear in mind that the increase in tax associated with each alternative is only intended to finance policies for the promotion of organic olive growing in the mountainous olive groves of Andalusia, which occupy 24% of the total area farmed with olives in Andalusia, and 8% of all agricultural land.

In T4, we combined both types of CTSs (i.e., those shown in T2 and T3) to test whether there is an effect of joint CTS presentation.

Between December 2012 and February 2013, 200 surveys were completed for each treatment by a specialised market research company using a random sampling procedure.[5] For each treatment, we recorded the time that each respondent spent completing the choice tasks and the percentage of inconsistent choices, determined using the procedure described below.

After completing the choice tasks, that is, *ex-post*, respondents' choices were screened by means of an iterative procedure in line with Colombo et al. (2016). In a follow-up question, we asked all respondents who chose a 'non-status quo' alternative at least once to state their maximum WTP for the best possible outcome according to non-monetary attribute levels. Based on expected monotonic preferences for non-monetary attributes, this outcome is characterised by the highest level of each non-monetary attribute. This payment card elicitation was

[4]Consequentiality is a necessary condition for incentive compatibility in DCEs.

[5]In total, 201 completed questionnaires were collected for T1.

presented in a format similar to alternatives in DCE choice tasks and is shown in Appendix A, online. Respondents were asked to state their maximum WTP using a payment ladder. The stated maximum WTP was then compared to the tax increase associated with the chosen alternatives to detect whether respondents' choices were inconsistent with the maximum WTP stated *ex post*. That is, we compared whether respondents had chosen an alternative in the choice tasks that represented worse environmental and social outcomes at a higher cost than the stated maximum WTP for the best outcome. In cases where this occurred, we asked respondents to review their decisions and allowed them to revise their choices if they wished to do so. The procedure is summarised in Appendix A, online. It provided information for an *ex-post* analysis of choices, in which revised choices replace initially 'inconsistent choices'. It is important to point out that to avoid forcing a reduction in WTP, we did not prompt or constrain respondents to alter their initial choice; rather, respondents could choose to retain their initial choice or revise it. We stored both initial responses (*ex-ante*) and responses following potential revision (*ex-post*), thus enabling an investigation of the effect of *ex-post* revision on WTP estimates. In the last part of the questionnaire, we gathered respondents' socio-economic data and other information about their current organic food consumption, as well as the reason for consuming or not consuming organic foods.

3 | ECONOMETRIC FRAMEWORK

According to random utility theory, utility respondent i obtained from alternative j at choice occasion t relies on a deterministic term $\boldsymbol{X}_{ijt}\boldsymbol{\beta}_i$ and a random term ε_{ijt} following a Gumbel distribution:

$$U_{ijt} = \boldsymbol{X}_{ijt}\boldsymbol{\beta}_i + \varepsilon_{ijt}, \tag{1}$$

where $\boldsymbol{X}_{ijt}$ is the vector of k attributes describing alternative j faced by respondent i at time occasion t, and $\boldsymbol{\beta}_i$ is the individual-specific vector of k preference parameters. In the MXL (McFadden and Train, 2000; Revelt and Train, 1998; Train, 2009), elements of $\boldsymbol{\beta}_i$ are modelled as random, following a parametric probability distribution the researcher selected *a priori*. MXL appears to be a state-of-the-art practice in the econometric analysis of discrete choice data. In addition, we apply the semi-parametric LML (Train, 2016) as an alternative way to model preference heterogeneity.

The MXL model accounts for preference or WTP heterogeneity, following a particular parametric distribution. The multivariate (parametric) distribution of these parameters in the sample is $\boldsymbol{\beta}_i \sim \boldsymbol{f}\,(\boldsymbol{b}, \boldsymbol{\Sigma})$, where $\boldsymbol{b}$ is a vector of sample means, and $\boldsymbol{\Sigma}$ is a variance–covariance matrix. A convenient way of accounting for preference differences associated with information treatments is $\boldsymbol{\beta}_i \sim \boldsymbol{f}\,\left(\boldsymbol{b} + z_i\boldsymbol{\delta}, \boldsymbol{\Sigma}\right)$, where z is a binary indicator for treatment effects and $\boldsymbol{\delta}$ is a vector of the estimated attribute-specific effects.

To facilitate interpretation of the results, we specify the model in WTP space (Train and Weeks, 2005) :

$$U_{ij}=\beta_i^m(X_{ijt}^m + \mathbf{X}_{ijt}^{-m}\boldsymbol{\beta}_i^{-m}) + \varepsilon_{ij}, \tag{2}$$

where X_{ij}^m is the monetary attribute with respect to which all marginal rates of substitution (WTP) are expressed, and $\boldsymbol{X}_{ij}^{-m}$ represents all other attributes. In this specification, parameter estimates $\left(\boldsymbol{\beta}_i^{-m}\right)$ can be readily interpreted as the marginal WTP for non-monetary attributes. Here, we can also define $\boldsymbol{\beta}_i^{-m} \sim \boldsymbol{f}\,\left(\boldsymbol{b}^{-m} + z_i\boldsymbol{\delta}^{-m}, \boldsymbol{\Sigma}^{-m}\right)$, which allows us to interpret $\boldsymbol{b}^{-m}$ as the mean WTP for a base treatment and $\boldsymbol{b}^{-m} + z_i\boldsymbol{\delta}^{-m}$ as the mean WTP for other treatments.

Estimation of the MXL requires calculation of the k dimensional integral for a likelihood function of individual i:

$$L_i = \int p\left(\boldsymbol{y}_i|\boldsymbol{X}_{ijt}, \beta_i^m, \boldsymbol{\beta}_i^{-m}\right) f\left(\beta_i^m, \boldsymbol{\beta}_i^{-m}|\Omega\right) d\left(\beta_i^m, \boldsymbol{\beta}_i^{-m}\right), \tag{3}$$

where $f\left(\beta_i^m, \boldsymbol{\beta}_i^{-m}|\Omega\right)$ is a density function of random parameters, whose distributions depend on the parameters to be estimated, Ω, and $p\left(\boldsymbol{y}_i|\boldsymbol{X}_{ijt}, \beta_i^m, \boldsymbol{\beta}_i^{-m}\right)$ is the conditional probability of making choices, $\boldsymbol{y}_i$, given by:

$$p\left(\boldsymbol{y}_i|\boldsymbol{X}_{ijt}, \beta_i^m, \boldsymbol{\beta}_i^{-m}\right) = \prod_t \left(\sum_j y_{ijt} \frac{\exp\left(\beta_i^m\left(X_{ijt}^m + \boldsymbol{X}_{ijt}^{-m}\boldsymbol{\beta}_i^{-m}\right)\right)}{\sum_l \exp\left(\beta_i^m\left(X_{ilt}^m + \boldsymbol{X}_{ilt}^{-m}\boldsymbol{\beta}_i^{-m}\right)\right)} \right). \tag{4}$$

As the analytical formula for the integral in Equation (3) is usually unknown, it must be approximated. Researchers typically employ the maximum simulated likelihood (MSL) method, in which R random draws from a distribution described by $f\left(\beta_i^m, \boldsymbol{\beta}_i^{-m}|\Omega\right)$ has to be generated for each individual. Using the draws, the integral can be approximated as:

$$L_i \approx \frac{1}{R} \sum_r p\left(\boldsymbol{y}_i|\boldsymbol{X}_{ijt}, \beta_{ir}^m, \boldsymbol{\beta}_{ir}^{-m}\right), \tag{5}$$

where the additional index r denotes the r th draw. To make the estimation more precise, we used 10,000 scrambled Sobol draws (Czajkowski and Budziński, 2015).

The LML model is a semi-parametric approach proposed by Train (2016), which allows for the estimation of the shape of the preference heterogeneity distribution without imposing restrictive assumptions regarding its parametric specification. Initial research suggests that it may be a promising new direction in discrete choice modelling, as it allows for the recovery of multimodal and asymmetric distributions (Franceschinis et al., 2017) and recovers induced means of respondents' WTP better than the standard MXL model (Bazzani et al., 2018).

The econometric specification of LML is not significantly different from that described above for the MXL. However, instead of assuming a parametric, continuous distribution for random parameters, such as $f\left(\beta_i^m, \boldsymbol{\beta}_i^{-m}|\Omega\right)$, the true (continuous) distribution is approximated using a discrete distribution. Specifically, we assume that the k th random parameter lies within some interval, $\left[LOW_k, UP_k\right]$, and we choose N points dividing this interval into $N-1$ smaller intervals of equal lengths. This creates a grid of N^K vectors of parameter values. The likelihood function is then given by:

$$L_i = \sum_{n=1}^{N^K} W\left(\beta_{in}^m, \boldsymbol{\beta}_{in}^{-m}|\boldsymbol{\alpha}, \lambda\right) p\left(\boldsymbol{y}_i|\boldsymbol{X}_{ijt}, \beta_{in}^m, \boldsymbol{\beta}_{in}^{-m}\right), \tag{6}$$

where $W\left(\beta_{in}^m, \boldsymbol{\beta}_{in}^{-m}|\boldsymbol{\alpha}, \lambda\right)$ is the probability of the vector of parameter values, $\beta_{in}^m, \boldsymbol{\beta}_{in}^{-m}$, which depends on the parameters to be estimated, $\boldsymbol{\alpha}, \boldsymbol{\lambda}$. The formula for this probability is given by a standard multinomial logit:

$$W(\beta_{in}^m, \boldsymbol{\beta}_{in}^{-m}|\boldsymbol{\alpha}, \lambda) = \frac{\exp(\boldsymbol{Z}(\beta_{in}^m, \boldsymbol{\beta}_{in}^{-m})\boldsymbol{\alpha} + \boldsymbol{V}(\beta_{in}^m, \boldsymbol{\beta}_{in}^{-m}, \boldsymbol{T}_i)\lambda)}{\sum_d \exp(\boldsymbol{Z}(\beta_{id}^m, \boldsymbol{\beta}_{id}^{-m})\boldsymbol{\alpha} + \boldsymbol{V}(\beta_{id}^m, \boldsymbol{\beta}_{id}^{-m}, \boldsymbol{T}_i)\lambda)}. \tag{7}$$

$\boldsymbol{Z}\left(\beta_{in}^{m}, \boldsymbol{\beta}_{in}^{-m}\right)$ in Equation (7) denotes some flexible transformation of the values of the random parameter vector. The transformations considered here are Legendre polynomials, step functions and four versions of splines (linear spline, cubic spline, piecewise cubic spline and piecewise cubic Hermite interpolating spline). To incorporate correlations between random parameters, we included first-order interactions between the elements of vectors $\left(\beta_{in}^{m}, \boldsymbol{\beta}_{in}^{-m}\right)$.

$\boldsymbol{V}\left(\beta_{in}^{m}, \boldsymbol{\beta}_{in}^{-m}, \boldsymbol{T}_i\right)$ in Equation (7) denotes some transformation of the values of the random parameter vector and individual-specific treatment, T_i. The incorporation of additional individual-specific explanatory variables into the LML framework has not yet attracted much attention. Approaches that have been considered to date include incorporating the additional interaction between socio-demographic/treatment variables and attributes directly into the utility function (Bansal et al., 2018b) or estimating a separate model for each value of the socio-demographic/treatment variable (Caputo et al., 2017). We consider both approaches to be suboptimal, as the former requires fixing the interaction parameter in the utility function, which reportedly makes estimation 20–40 times longer, and the latter may be infeasible if socio-demographic/treatment variables take multiple values. Here, we incorporate socio-demographic/treatment variables as an interaction with random parameter values, namely $\boldsymbol{V}\left(\beta_{in}^{m}, \boldsymbol{\beta}_{in}^{-m}, \boldsymbol{T}_i\right) = \left(\boldsymbol{T}_i\beta_{in}^{m}, \boldsymbol{T}_i\boldsymbol{\beta}_{in}^{-m}\right)$. Obviously, more complex functions can be defined, such as interaction with polynomials of $\beta_{in}^{m}, \boldsymbol{\beta}_{in}^{-m}$ of a higher order than one, but this also requires estimation of a greater number of coefficients.

Similar to the MXL case, model estimation can be performed using the MSL method. We used R random draws from the grid (each point was drawn with the same probability) for each individual to approximate the likelihood function:

$$L_i \approx \sum_{r=1}^{R} W(\beta_{ir}^{m}, \boldsymbol{\beta}_{ir}^{-m}|\boldsymbol{\alpha}, \lambda)p(\boldsymbol{y}_i|\boldsymbol{X}_{ijt}, \beta_{ir}^{m}, \boldsymbol{\beta}_{ir}^{-m}). \tag{8}$$

As in the case of MXL, we used scrambled Sobol draws to make the estimation more efficient. The approximation in Equation (8) can be used to calculate the mean WTP, as a sum of R random draws from the grid, weighted by the estimated probability that they will occur in the population, $W\left(\beta_{ir}^{m}, \boldsymbol{\beta}_{ir}^{-m}|\boldsymbol{\alpha}, \boldsymbol{\lambda}\right)$:

$$MWTP \approx \sum_{r=1}^{R} W(\beta_{ir}^{m}, \boldsymbol{\beta}_{ir}^{-m}|\boldsymbol{\alpha}, \lambda)\boldsymbol{\beta}_{ir}^{-m}. \tag{9}$$

Working with the LML model requires the selection of an appropriate specification. Multiple options are available for the specification of the $\boldsymbol{Z}\left(\beta_{in}^{m}, \boldsymbol{\beta}_{in}^{-m}\right)$ function. Most existing studies used information criteria to guide specification choice. We also employed this approach.[6] Finally, model specification requires selecting the values of the random distribution bounds. Most studies used the estimates from the MXL model, with bounds defined

[6]Similar to Bazzani et al. (2018), we find that the LML does not provide an improvement over the MXL model if the Bayesian information criterion (BIC) is used as a basis for comparison. Bansal et al. (2018a) noted that this is the most likely in small samples and recommended considering significant changes in histograms of parameter distributions and choosing a minimal number of parameters so that any additional parameters will not change the shape of the distribution substantially. We found this approach hard to implement in practice, as it was difficult to assess whether observed changes in shape should be considered significant. Additionally, following this approach would likely exponentially increase the number of model specifications to consider. As a result, in what follows, we used the Akaike information criterion (AIC) to compare the fit of different specifications.

as two standard deviations above and below the mean. Train (2016) was the first to use this approach, and to the best of our knowledge, only Caputo et al. (2017) experimented with different settings by taking three standard deviations or extending parameter bounds based on visual inspection. The results they obtained are mixed; for some specifications of $\boldsymbol{Z}\left(\beta_{in}^{m}, \boldsymbol{\beta}_{in}^{-m}\right)$, extending parameter bounds increases the model fit, but for other specifications, it decreases the model fit. We also test the sensitivity of model performance depending on the range of bounds, which are defined with reference to the parameters of an MXL model without correlation.

4 | RESULTS

The overall sample is representative of the Andalusian population with respect to gender (chi-squared = 0.12; p-value = 0.73). However, it differs in terms of age, education and income, with younger (chi squared = 2399, 0; p-value = 0.00) and highly educated citizens (chi squared = 1125.5; p-value = 0.00) being over-represented. The sample is representative in terms of lower income ranges, but it is not representative in terms of medium- and higher-income respondents, who are over- and under-represented, respectively (chi squared = 52.5; p-value = 0.00). These differences were maintained across the four treatments, which were all different relative to the general population with respect to the dimensions described above. That is, there were no significant differences in socio-economic characteristics across treatments. Treatments do not affect the response time related to choosing the preferred option in the choice tasks. In all treatments, respondents became faster at completing the choice tasks as they moved through the sequence of choices, as observed previously in other studies (Bonsall and Lythgoe, 2009; Carlsson et al., 2012). The reduction in completion times was particularly pronounced between the first and second choice tasks. Furthermore, treatments showed a similar percentage of 'inconsistent' choices, with values close to 22% for *ex-ante* choices and 12% for *ex-post* choices.

Table 2 presents respondents' estimated WTP for the attribute levels elicited for each of the four CTS treatments using *ex-ante* data (the initial choices, before respondents were given a chance to revise inconsistencies with the follow-up maximum WTP question). The models use dummy coding for all attributes and assume that their parameters are normally distributed, except for cost, entering models on a continuous scale, and the associated (negative) cost parameter, which was assumed to be log-normally distributed. The left panel of Table 2 presents the mean WTP estimates based on the MXL model. The right panel of Table 2 presents the equivalent results based on the best fitting LML model. Results presented in Table 2 refer to models that combine observations from all treatments ('All treatments jointly') or allow for treatment-specific WTPs.

To find the best-fitting LML model specification, we first applied a grid-search procedure to examine the sensitivity of the estimated log-likelihood at convergence to the specification of the parameter bounds. We found that using bounds specified as 1.5–2.5 MXL-based standard deviations below and above the mean resulted in relatively similar results, with the optimum identified at 1.8. This supports the use of the rule of thumb of approximately two standard deviations below and above the mean (Train, 2016). However, it must be noted that this approach did not differentiate the bounds for each parameter, implement asymmetric bounds, or generate other reference points for determining bounds (i.e., not based on the results of the MXL model without correlations). The results for the effects of selecting different bound ranges are provided in Appendix B, online. Next, we compared the performance of various LML model specifications (asymptotic normal, polynomial, step function, and four types of spline functions of 2–10 degrees) with and without correlated parameters. A comparison of the fit of various LML model specifications is available in Appendix C, online. Based on the AIC, we selected the 8-knot piece-wise cubic spline

TABLE 2 Mean WTP for policy attributes estimated using joint and treatment-specific data and the MXL and LML models using *ex-ante* data [EUR/year]

	MXL					LML				
	All treatments jointly	Control: standard remainders	Treatment 2: HB cheap talk	Treatment 3: Scale of the project reminder	Treatment 4: HB cheap talk and scale of the project reminder	All treatments jointly	Control: standard remainders	Treatment 2: HB cheap talk	Treatment 3: scale of the project reminder	Treatment 4: HB cheap talk and scale of the project reminder
Status quo (alternative specific constant)	31.90*** (3.55)	27.60 (4.90)	27.63 (4.96)	37.73 (5.67)	37.12 (6.54)	18.74*** (1.32)	18.30 (2.36)	16.53 (2.35)	23.13* (2.28)	20.72 (2.07)
Tackling climate change—medium (vs. low)	15.22*** (1.60)	16.34 (2.36)	11.35 (2.22)	16.03 (2.88)	17.71 (2.62)	13.64*** (0.12)	13.55 (0.21)	13.78 (0.22)	13.30 (0.20)	14.14** (0.21)
Tackling climate change—high (vs. low)	19.88*** (1.70)	18.82 (2.50)	16.91 (2.26)	21.45 (2.77)	21.71 (2.64)	17.59*** (0.37)	17.87 (0.65)	16.84 (0.56)	17.91 (0.57)	17.41 (0.63)
Biodiversity—medium (vs. low)	15.16*** (1.57)	17.77 (2.56)	14.82 (2.33)	15.46 (2.97)	12.14 (2.78)	13.15*** (0.28)	13.16 (0.47)	14.19 (0.53)	13.64 (0.48)	13.11 (0.62)
Biodiversity—high (vs. low)	21.10*** (1.71)	23.14 (2.57)	19.27 (2.37)	21.10 (2.79)	20.76 (2.69)	18.98*** (0.41)	19.92 (0.64)	18.22* (0.78)	18.03** (0.69)	19.68 (0.80)
Risk of pollution of water resources—moderate (vs. high)	18.67*** (1.62)	17.08 (2.58)	17.40 (2.33)	23.25 (3.01)	18.40 (2.75)	16.04*** (0.08)	16.09 (0.15)	16.05 (0.15)	16.20 (0.17)	15.99 (0.16)
Risk of pollution of water resources—low (vs. high)	28.10*** (2.09)	28.19 (2.94)	24.56 (2.91)	32.51 (3.55)	27.86 (3.26)	26.29*** (0.22)	26.54 (0.39)	26.25 (0.38)	26.32 (0.39)	25.92 (0.40)
Soil erosion—moderate (vs. high)	12.05*** (1.57)	10.32 (2.49)	11.35 (2.44)	13.20 (2.79)	13.50 (2.67)	10.89*** (0.02)	10.85 (0.04)	10.89 (0.03)	10.89 (0.04)	10.93* (0.03)
Soil erosion—low (vs. high)	17.84*** (1.57)	16.79 (2.52)	17.79 (2.21)	19.71 (2.74)	17.53 (2.64)	15.28*** (0.22)	15.12 (0.41)	15.45 (0.47)	16.34** (0.46)	14.61 (0.36)
Agricultural employment—5% increase (vs. no change)	15.50*** (1.63)	17.27 (2.76)	13.56 (2.29)	14.68 (2.96)	17.61 (2.72)	15.80*** (0.03)	15.83 (0.05)	15.75 (0.05)	15.80 (0.05)	15.81 (0.05)

(Continues)

TABLE 2 (Continued)

	MXL					LML				
	All treatments jointly	**Control: standard remainders**	**Treatment 2: HB cheap talk**	**Treatment 3: Scale of the project reminder**	**Treatment 4: HB cheap talk and scale of the project reminder**	**All treatments jointly**	**Control: standard remainders**	**Treatment 2: HB cheap talk**	**Treatment 3: scale of the project reminder**	**Treatment 4: HB cheap talk and scale of the project reminder**
Agricultural employment—10% increase (vs. no change)	27.24*** (1.98)	32.56 (3.46)	21.67** (2.84)	24.26* (3.41)	30.81 (3.26)	24.42*** (0.75)	24.82 (1.17)	24.05 (1.24)	22.51 (1.19)	25.64 (1.11)
Model diagnostics										
LL at convergence	−3939.19	−3921.03				−3784.35	−3753.14			
LL at constant(s) only	−5279.91	−5279.91				−5279.91	−5279.91			
McFadden's pseudo-R^2	0.2539	0.2574				0.2833	0.2892			
Ben-Akiva-Lerman's pseudo-R^2	0.4609	0.4626				0.4808	0.4846			
AIC/n	1.6767	1.6842				1.6473	1.6492			
BIC/n	1.7980	1.8540				1.8818	1.9323			
n (observations)	4806	4806				4806	4806			
r (respondents)	801	801				801	801			
k (parameters)	90	126				174	210			

Notes: *, **, and *** indicate the statistical significance (Wald test) of the difference in WTP between the control treatment (standard remainders) and the other treatments (additional measures aimed at reducing hypothetical bias) at the 0.1, 0.05 and 0.01 levels, respectively. For 'All treatments jointly', asterisks indicate statistical significance with respect to 0. Standard errors are shown in parentheses.

as the best-fitting specification of the LML model.[7] It is worth mentioning that for the models with correlated parameters, based on the Bayesian information criterion, which is more restrictive in terms of penalising models for the number of parameters, none of the LML models outperformed the MXL specification in terms of model fit.

The first thing to note about the results presented in Table 2 is that there are some differences between the mean WTP estimates implied by the MXL and LML models. These differences are within approximately 10% of the WTP derived from the MXL model and are not statistically different according to *z*-tests, except for the status quo, for which WTP resulting from the LML model was 40% lower.[8]

Looking at the differences in WTP associated with treatments, we find that in almost all cases, estimates of mean WTP resulting from T2–4 are not significantly different from WTP estimates inferred from the control treatment (T1). In the case of the MXL model, the only significant difference was observed for a 10% increase in agricultural employment for T2 and T3. Irrespective of the statistical significance of the differences, there is no consistent trend of lower mean WTP estimates arising from single or joint treatment with CTSs. For the LML model results, significant differences were observed in more cases than in the MXL model results. This is primarily due to the lower standard errors of the parameter estimates of the LML model relative to the MXL model. Significant differences represent either an increase (e.g., soil erosion—high: T3) or a decrease (e.g., biodiversity—high: T2 and T3) in the WTP estimates of attribute levels.

The analysis of WTP after respondents' review of their 'inconsistent' choices (*ex-post*) provides similar results. The results (presented in Table A3, online) show that in most cases, the additional cheap talk and reminders about the extent of the project (scale) used in T2–4 did not result in statistically different WTP estimates for the MXL model, relative to T1, which only used standard budget constraint reminders. Again, a greater number of significant differences were found for the LML model results, but there is no uniform trend of lower WTP estimates for T2–4 relative to T1 across all attributes. Furthermore, the share of inconsistent choices was not statistically different between any of the four treatments, revealing that the use of CTSs does not affect the degree of choice inconsistency of respondents. Overall, we conclude that the inclusion of CTSs had a limited effect on WTP estimates for both respondents' initial and revised choices.

However, we find that allowing respondents to revise their choices leads to significant reductions in WTP. This result is illustrated in Figure 2, which presents estimates of mean WTP resulting from MXL and LML models for *ex-ante* (unrevised) and *ex-post* (revised) data along with the 95% confidence interval. According to *z*-tests, the mean WTP estimates for all attributes are statistically different at the 5% significance level for MXL and LML models apart from the ASC and the medium level of the biodiversity attribute in the LML model. Allowing respondents to reconsider their choices leads, on average, across all attributes, to a WTP decrease of 43% (MXL model) or 33% (LML model). This effect is considerable, given that respondents were neither prompted nor constrained to change their initial choices in the revision process. The percentage of 'inconsistent' choices dropped from 21.8% before the revision to 12.5% after the revision, indicating that a significant part of the sample opted to retain their initial choices.

The procedure used to identify choice consistency allows us to investigate the potential reasons underlying the decisions to revise or not to revise previous choices. This can provide supporting information for understanding reasons for choice inconsistencies and associated bias. The analysis of inconsistent choices revealed that of 801 individuals, 342 (43%) responded

[7]The models presented here were estimated using a DCE package developed in Matlab that is available at https://github.com/czaj/DCE. The code and data for estimating the specific models presented in this study, as well as full and supplementary results, are available from https://czaj.org/research/supplementary-materials.

[8]As an aside, we found that the status quo parameter estimate was relatively the least stable across various LML model specifications, possibly because of the least variability in experimental design. For the rest of the parameter estimates, the standard errors resulting from the LML model are substantially lower than the standard errors associated with the MXL-based estimates.

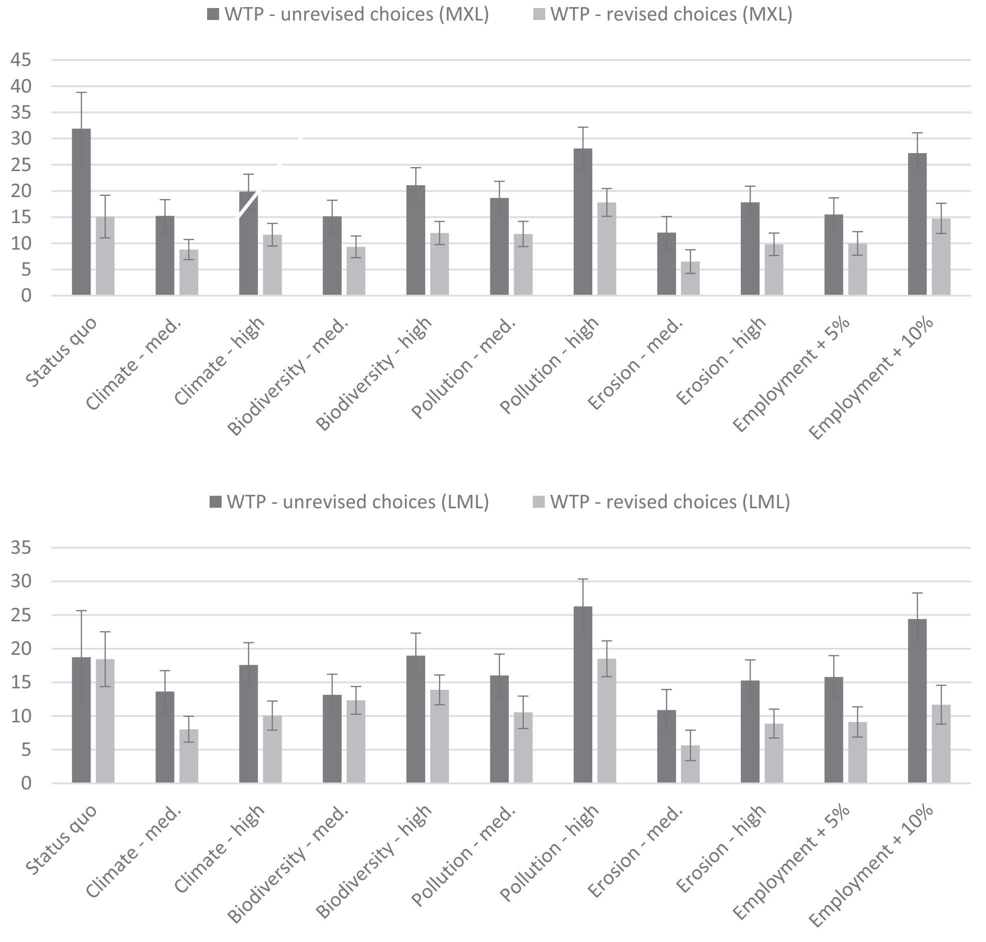


FIGURE 2 Comparison of WTP means between the estimated models

consistently to all choice tasks, while 459 respondents were found to have at least one inconsistent choice (one: 170, two: 116, three: 94, four: 42, five: 28, six: 9).

There are several explanations for why respondents make inconsistent choices. The first group of causes includes reasons associated with what we refer to as 'unpremeditated' inconsistent choices. These are choices where respondents unintentionally select choice alternatives with higher associated costs relative to their 'maximum' WTP.[9] We attribute this behaviour to individuals who, once faced with their inconsistencies, revised all of them to make them consistent or who, in line with the soft cutoffs approach (Bush et al., 2009; Swait, 2001), made a

[9]We refer to 'maximum' WTP in quotation marks acknowledging that, as explained in the study design section, we do not have data on real WTP that could be used as a benchmark. Consequently, we implicitly assume that the WTP stated on the payment card is respondents' maximum WTP. We acknowledge that this may attract criticism, given that the payment card has not been theoretically shown to be fully incentive compatible (Vossler and Holladay, 2018). However, this format maintains choice set independence and one-to-one matching between the project identified and the policies that might be implemented. There is insufficient empirical evidence from the field to understand whether the format generates demand revealing responses if accompanied by perceived consequentiality.

'small' violation of the stated price cutoffs.[10] Inconsistent choices due to 'unpremeditated' HB may have occurred because the hypothetical context is less demanding than a real payment context, increasing the likelihood of positive responses. This is supported by empirical evidence on HB from neuroimaging studies that found that brain activity showed that hypothetical and real choices both activated common areas of the brain, but that the activation strength was higher in the real choice condition (Kang et al., 2011). This is also observed in the case of moral choices, where common neural networks are observed for both real and hypothetical conditions. However, there is a distinct circuitry of activity specific to each condition (Feldman-Hall et al., 2012).

In the case of public goods, 'unpremeditated' HB may also be linked to respondents' desire to comply with social norms and behaviours, given that people derive utility from portraying a positive self-image (Johansson-Stenman and Svedsäter, 2012; Morrison and Brown, 2009). This may have led respondents to focus on the non-monetary attributes of the choice alternatives, all linked with the provision of positive environmental and social outcomes, without trading them off with the associated price. In addition, the multi-attribute design of DCE may reduce the importance of price relative to other attributes and thus distract respondents' attention from it. Similarly, respondents facing a choice set for the first time may interpret it differently than subsequent choice sets because of institutional learning (Meyerhoff and Glenk, 2015). In this case, the possibility of respondents reconsidering their choice is a way of correcting unintended errors.

Another possibility is the existence of uncertainty regarding WTP for the studied good. In this case, the maximum WTP should not be considered as a fixed amount but rather as a value with a variance that is proportional to the degree of fuzziness or uncertainty (Sun and Van Kooten, 2009). Carlsson et al. (2012) observed that in this case, individual choice in a public good context can be expected to diverge significantly from standard utility theory predictions if preferences are well defined. It is also possible that the two WTP elicitation methods employed in this study disclosed different values. Roe et al. (1996) and Salensminde (2003) observed that DCE, compared to open-ended CV, tends to capture respondents' relative rather than absolute valuation, which is more in line with their budget constraints. Vossler and Holladay (2018) observed that the payment card format used to disclose the maximum WTP may underestimate the values. Thus, it may be expected that the values obtained from the payment card CV would be lower than those derived through DCE.

In the second group of potential reasons underlying inconsistent choices, related to SB, we include explanations related to the voluntary choice of individuals who *deliberately* violate their maximum stated WTP to strategically influence policy outcomes. Such behaviour is consistent with respondents who, when faced with inconsistent choices relative to the stated maximum WTP, decided to maintain the original choices, even if the violation of the price cutoffs is above an acceptable limit. The cutoff threshold value is an empirical issue that must be set according to the study's specific conditions and the available information (Bush et al., 2009). In general, when analysts expect larger uncertainties, higher values should be used. In this study, acknowledging that the good under study is unfamiliar to respondents and that the threshold value affects the number of individuals that are considered to be unintentionally or strategically inconsistent, we employ four different values of the acceptable limit for the price cutoff violation: 0%, 30%, 70% and 100%. The 0% value only considers consistent individuals and treats all the other individuals as inconsistent,

[10]Small violations of the stated price cutoffs occur when the difference between the cost of the chosen alternative and the stated maximum WTP is within the analyst's accepted violation limit. For instance, the violation of a respondent who stated €50 as maximum WTP and chose an alternative whose cost is €51 can be considered within the acceptable limit. As pointed out further on, several limit thresholds are used in this paper.

either because their choices were strategically made or due to any other reasons. Values other than 0% acknowledge a certain degree of uncertainty in the stated maximum WTP before considering an individual as inconsistent. This behavioural pattern applies to 277, 197, 103 and 94 respondents for the 0%, 30%, 70% and 100% price cut-off limits, respectively, which constitute a substantial share of the sample.

We acknowledge that there might be other reasons for respondents to retain their inconsistent choices. Fatigue arising from the repetitive nature of choice tasks, the desire to speed up progress through the survey, indifference to changing anything, and a feeling of uncertainty about what to do or whether to contradict earlier responses may all be possible causes for not changing inconsistent choices. Nevertheless, we attribute choice inconsistency to HB or SB in this study for the following reasons. First, we did not find any differences in the time that inconsistent respondents needed to complete the survey relative to consistent respondents.[11] This suggests that inconsistent respondents did not pay different levels of attention to the choice task or tend to complete the survey as quickly as possible.

We found that the group of inconsistent respondents comprised consumers who frequently purchase organic food and stated that the main reason for doing so is because organic food consumption contributes to protecting the environment.[12] This supports a plausible explanation that these individuals deliberately chose the policy alternatives that provide the best environmental outcomes irrespective of the associated cost, with the aim of influencing the study's results. This group of respondents shares characteristics with environmentalists, who often complain about the low level of attention that policy in general and agricultural policy in particular pay to the environment. Thus, these respondents may have perceived the possibility of influencing agricultural policy through the survey and therefore opted to vote strategically through their responses. In this context, the use of a non-incentive compatible choice mechanism (repeated choice from choice set containing three options) may also have encouraged the respondents' strategic behaviour.

To illustrate the effect on mean WTP estimates, we estimated additional models that included all the consistent individuals and excluded respondents who made what might be a strategic or an inconsistent choice due to any other reason. Given the different treatments' lack of statistical significance, we estimated a joint model that included all treatment observations (Table 3).

The results of the MXL model indicate that mean WTP estimates are significantly affected by the inconsistent choices of the individuals who were identified to behave strategically; however, this is only the case if we do not allow any degree of uncertainty in the respondents' stated maximum WTP. If respondents' stated maximum WTP is allowed to be up to 30% lower than the price associated with the alternative selected in choice tasks, WTP differences are only marginally significant for one of the attributes. No significant differences were observed if the analyst was willing to accept an uncertainty threshold of 70%. The results of the LML model also indicate significant differences in the mean WTP. However, in this case, the significance of the differences does not necessarily diminish with a higher threshold of accepted uncertainty. This is possibly a result of the very high sensitivity of the standard errors derived from the highly parametrised LML model for constrained samples. In our case, this unreliability precludes us from drawing clear conclusions with respect to whether removing inconsistent choices (at an assumed uncertainty level) significantly influences the observed WTP estimates.

[11]We employed Mann-Whitney tests for the comparison and found no differences in the time taken to complete the survey between the respondents who retained their initial choices compared to those who revised them.

[12]Using chi-squared tests, we observed that the group of consistent individuals differs statistically, relative to the inconsistent group, in the frequency of organic food consumption and in the stated motivation. Inconsistent individuals declared that the main reason for organic food consumption is related to protecting the environment.

TABLE 3 MXL mean WTP for policy attributes estimated using joint *ex-post* data and different price cutoff limits [EUR/year]

	MXL				LML			
	Price cutoff limit (0%)	**Price cutoff limit (30%)**	**Price cutoff limit (70%)**	**Price cutoff limit (100%)**	**Price cutoff limit (0%)**	**Price cutoff limit (30%)**	**Price cutoff limit (70%)**	**Price cutoff limit (100%)**
Status quo (alternative specific constant)	11.98 (1.98)	14.96 (2.19)	15.03 (2.17)	14.44 (2.14)	19.8 (3.15)	15.64 (1.26)	10.04*** (1.78)	12.65* (1.52)
Tackling climate change—medium (vs. low)	5.30* (1.03)	7.53 (1.09)	7.99 (1.02)	8.00 (1.07)	5.14*** (0.37)	6.97*** (0.24)	8.39 (0.25)	7.38* (0.22)
Tackling climate change—high (vs. low)	7.54* (1.09)	9.79 (1.22)	10.80 (1.13)	10.73 (1.15)	7.86* (1.43)	11.56* (0.59)	10.41 (0.6)	10.27 (0.45)
Biodiversity—medium (vs. low)	6.12* (1.08)	7.72 (1.16)	8.43 (1.06)	8.05 (1.06)	10.67* (0.6)	9.24*** (0.31)	9.43*** (0.24)	10.43*** (0.21)
Biodiversity—high (vs. low)	8.09* (1.06)	9.88 (1.14)	10.64 (1.09)	10.51 (1.11)	10.61*** (1.09)	12.63 (0.41)	11.44*** (0.55)	13.17 (0.41)
Risk of pollution of water resources—moderate (vs. high)	7.64* (1.07)	10.04 (1.13)	10.87 (1.21)	10.88 (1.21)	11.12 (0.14)	11.07*** (0.05)	11.27*** (0.07)	11.08*** (0.06)
Risk of pollution of water resources—low (vs. high)	11.66*** (1.30)	14.28* (1.48)	15.90 (1.39)	15.97 (1.40)	16.86 (0.86)	18.48 (0.3)	19.07* (0.43)	17.13* (0.42)
Soil erosion—moderate (vs. high)	4.30 (1.21)	5.75 (1.24)	6.01 (1.17)	5.30 (1.18)	6.09 (0.27)	6.21* (0.12)	4.95 (0.12)	5.77 (0.11)
Soil erosion—low (vs. high)	5.36*** (1.08)	7.84 (1.23)	8.77 (1.11)	8.28 (1.08)	8.07 (0.76)	7.49* (0.36)	8.76 (0.35)	7.47* (0.31)
Agricultural employment—5% increase (vs. no change)	6.68* (1.02)	8.54 (1.17)	8.99 (1.09)	9.06 (1.08)	9.12 (0.01)	9.13 (0.01)	9.14 (0.01)	9.15*** (0.01)
Agricultural employment—10% increase (vs. no change)	9.22* (1.44)	12.74 (1.60)	13.22 (1.44)	13.61 (1.38)	14.75 (1.59)	12.08 (0.72)	11.82 (1.01)	13.07 (0.82)
Model diagnostics								
LL at convergence	−2208.40	−2685.12	−3167.47	−3219.50	−2169.92	−2639.42	−3132.47	−3196.07
LL at constant(s) only	−3268.07	−3871.29	−4525.07	−4587.20	−3268.07	−3871.29	−4525.07	−4587.20

(Continues)

TABLE 3 (Continued)

	MXL				LML			
	Price cutoff limit (0%)	**Price cutoff limit (30%)**	**Price cutoff limit (70%)**	**Price cutoff limit (100%)**	**Price cutoff limit (0%)**	**Price cutoff limit (30%)**	**Price cutoff limit (70%)**	**Price cutoff limit (100%)**
McFadden's pseudo-R^2	0.3242	0.3064	0.3000	0.2982	0.3360	0.3182	0.3078	0.3033
Ben-Akiva-Lerman's pseudo-R^2	0.5242	0.5022	0.4928	0.4912	0.5398	0.5164	0.5035	0.4996
AIC/n	1.4621	1.5315	1.5556	1.5604	1.4910	1.5527	1.5790	1.5889
BIC/n	1.6354	1.6854	1.6919	1.6951	1.8261	1.8501	1.8424	1.8495
n (observations)	3144	3624	4188	4242	3144	3624	4188	4242
r (respondents)	524	604	698	707	524	604	698	707
k (parameters)	90	90	90	90	174	174	174	174

Notes: *, **, and *** indicate the statistical significance (Wald test) of the difference in WTP between the model estimated with all observations after the revision of inconsistent choices and the models estimated by filtering the respondents identified as 'strategic', at the 0.1, 0.05 and 0.01 levels, respectively. Standard errors are shown in parentheses.

Finally, we investigated if the effects of the *ex-post* treatment were more pronounced for the ASC associated with the status quo option than for the other parameters. As pointed out by an anonymous reviewer, SB might manifest itself through non-demand revealing choices of the status quo option, which in our models would be indicated by significant and positive parameters of the ASC. A comparison of estimates of the ASC in Table 2 (*ex-ante*) and Appendix D, online (*ex-post*) shows that values of the ASC are significant and positive, yet significantly smaller in magnitude for *ex-post* data. This might at first glance be interpreted as evidence of SB reduction through *ex-post* treatment. However, because models in Table 2 and Appendix D are estimated in WTP-space, the magnitude of the ASC parameters may indicate a difference in propensity to choose the status quo, a difference in sensitivity to cost-scale parameter, or a difference in both. To shed light on this, we estimated an additional preference-space model, which allows for separating the effects for means of the ASC, individual attributes and the cost parameter. To be able to compare coefficients between different models estimated in preference-space we combined *ex-ante* and *ex-post* treated data, essentially doubling the observations while dropping the inconsistent observations in the doubled part of the new dataset.[13]

The results of the model presented in Table 4 show that the effect of *ex-post* treatment is only observed for the cost parameter—the interactions are not significant for any of the other attribute means, including the ASC. This means that the mean WTP differences between untreated and *ex-post* treated coefficients were caused by the differences in respondents' sensitivity to cost and are thus not related to differences in SB related to status quo choice. As an aside, the results shown in Table 4 also suggest that differences in WTP for non-monetary attributes are also due to differences in estimated sensitivity to cost rather than preferences for programme attributes.

5 | DISCUSSION AND CONCLUSIONS

The hypothetical nature of DCE lies at the heart of the controversies about the reliability and validity of WTP estimates for non-market goods. This has given rise to methods aimed at testing and reducing hypothetical bias (HB). The results of our study add to this literature by observing that HB has significant implications for WTP estimates and demonstrating the effects of two methods to detect and reduce it.

The use of cheap talk scripts (CTSs) as an *ex-ante* tool to reduce HB has had a limited effect on WTP estimates for the environmental goods valued in this study. This was observed despite having displayed the CTSs for 30 s on the respondents' screen without the option to proceed and despite the fact that the CTSs were augmented with a single opt-out reminder. There may be several reasons for this result. First, the CTS effect may have lost effectiveness as respondents progressed through a series of choice tasks, as Ladenburg and Olsen (2014) suggested. When moving beyond the first choice tasks, respondents' attention to the message of the CTSs may have faded, especially considering the implicit difficulty of the choice tasks due to both unfamiliarity with the public goods and with answering to choice tasks. This finding can be interpreted as if CTSs have not affected the payment consequentiality of the survey, which in turn did not affect choices. The online format of the survey may also contribute to this result. As observed by Penn and Hu (2019), CTSs tend

[13]The effect of the treatment in our model is identified by the interactions of dummy-coded '*ex-post* treatment' with the mean of the preference distribution of each attribute. We do this, while accounting for the possible scale parameter differences between untreated and *ex-post* treated data; note that the identification of the scale parameter is based on the variations of random parameter distributions around the means and the correlations, and that the estimated scale parameter controls for the possible scale-driven differences in the means of parameters.

TABLE 4 The results of the MXL model in preference-space (combined untreated and treated data)

	Means	Interaction–Price cutoff limit (0%)	Standard deviations
Status quo (alternative specific constant)	1.6170*** (0.2585)	0.2618 (0.3669)	4.2809*** (0.3087)
Tackling climate change—medium (vs. low)	1.2666*** (0.1342)	0.0753 (0.1784)	1.1258*** (0.1720)
Tackling climate change—high (vs. low)	1.5574*** (0.1521)	0.0297 (0.1984)	1.3041*** (0.1962)
Biodiversity—medium (vs. low)	1.2911*** (0.1477)	0.0664 (0.1906)	0.8473*** (0.1916)
Biodiversity—high (vs. low)	1.5029*** (0.1471)	0.0494 (0.1896)	1.4221*** (0.2050)
Risk of pollution of water resources—moderate (vs. high)	1.4218*** (0.1351)	0.0996 (0.1884)	1.5307*** (0.2042)
Risk of pollution of water resources—low (vs. high)	2.0932*** (0.1636)	0.1497 (0.2141)	1.9004*** (0.2160)
Soil erosion—moderate (vs. high)	0.5104*** (0.1397)	−0.0699 (0.1888)	1.2513*** (0.2231)
Soil erosion—low (vs. high)	0.8143*** (0.1295)	−0.2223 (0.1810)	1.3389*** (0.1824)
Agricultural employment—5% increase (vs. no change)	1.1816*** (0.1306)	0.0215 (0.1834)	1.2186*** (0.2110)
Agricultural employment—10% increase (vs. no change)	1.9848*** (0.1876)	−0.0793 (0.2472)	2.8675*** (0.7082)
Cost (10 EUR/year)	0.2510*** (0.0758)	0.5656*** (0.1077)	1.3535*** (0.0934)
Model diagnostics			
LL at convergence	−5824.00		
LL at constant(s) only	−8555.41		
McFadden's pseudo-R^2	0.3193		
Ben-Akiva-Lerman's pseudo-R^2	0.5047		
AIC/n	1.4908		
BIC/n	1.5804		
n (observations)	7950		
r (respondents)	1325		
k (parameters)	102		

Notes: *, **, *** indicate statistical significance (with respect to 0) at the 0.1, 0.05 and 0.01 level, respectively. Standard errors provided in parentheses.

to be relatively ineffective for internet-based surveys compared to mail surveys. Second, it may be that the use of a budget reminder was sufficient to make respondents consider substitutes to the proposed changes and their budget constraints, so that the CTSs did not add any additional prompt regarding the careful consideration of the use of their budget. Third, the definition of the cost vector may have included tax amounts that were too low to choke off demand. The values were based on in-person qualitative pretests (focus groups), which may differ from online settings. Defining the cost vector levels is still a pending issue in the DCE literature, especially considering its importance for deriving WTP estimates (Glenk et al., 2019).

The CTS related to the scale of the project (T3) also proved to be relatively ineffective. This is remarkable given that respondents passed an internal scope test (i.e., WTP was found to be equal or higher for greater benefits in all attributes). However, as Rolfe and Wang (2011) pointed out, a project's scale and scope are typically intertwined aspects that respondents often struggle to separate. The ineffectiveness of a CTS related to project scale while internal scope is demonstrated may be attributed to respondents expressing relative values that result in internally consistent choices that may, however, appear inconsistent if compared across treatments. This is in line with the idea of coherent arbitrariness (Ariely et al., 2003). Proximity and loyalty effects can also be present for locally iconic crops such as olives (Granado-Díaz et al., 2020). Respondents may attach higher values to goods provided within nearer proximity (local goods) than those provided further away, especially for environmental goods that may inspire a sense of identification in people (Faccioli et al., 2020). Thus, independent of its scale (extent of provision), a good that elicits an emotional response may be highly valued (LaRiviere et al., 2014). Alternatively, respondents may simply be insensitive to the project scale, as other authors have previously observed (Rolfe and Windle, 2003).

Joint presentation of the reminders (T4) also proved ineffective. Therefore, using multiple CTSs may simply add complexity to the survey without providing additional benefits. This outcome is in line with Varela et al. (2014), who found that the use of multiple *ex-ante* mitigation strategies does not impact WTP estimates. However, it contrasts with results from other studies that found that the use of multiple *ex-ante* mitigation strategies can reduce HB (Jacquemet et al., 2013; Ladenburg and Olsen, 2014). All of these studies tested different *ex-ante* measures. Varela et al. (2014) employed CTS plus a single opt-out reminder, Ladenburg and Olsen (2014) compared the effect of CTS with multiple opt-out reminders, and Jacquemet et al. (2013) combined CTS and a solemn oath. In our study, we employed two different CTSs and a single opt-out reminder. In summary, the effect of CTS seems to depend on the way it is designed for each specific condition.

Approaches that revise respondents' choices after the choice task (*ex-post*) appear better positioned to reduce HB. Estimates of WTP decrease, on average, by 43% in the case of the MXL model and 33% in the LML model. This effect is statistically significant. It illustrates the quantitative importance of HB in the estimated WTP values. The adaption of mitigation measures to reduce HB can reduce the potential for inefficient allocation of resources in the design of public policies of non-marketed goods. Thus far, *ex-post* approaches to mitigate HB are far less widespread than *ex-ante* approaches. The results of this study clearly reveal that research on *ex-post* mitigation instruments, at least in public good contexts, deserves greater attention. As a desirable feature of the *ex-post* mitigation instrument applied here, we demonstrate that it is effective through affecting cost sensitivity of respondents in the absence of statistically significant effects on the tendency to choose the status quo (ASC) and non-monetary attribute preferences.

The results also raise the question of why the *ex-post* mitigation measure used outperformed *ex-ante* mitigation measures, a result that has been observed in previous research (Champ et al., 2009; Morrison and Brown, 2009). In our opinion, there are several reasons for this. First, *ex-ante* approaches do not reveal any information about respondents' attention to, and understanding of, the information provided, unless specifically inquired. Thus, the use of CTSs should be accompanied by questions that allow the analyst to appreciate respondents' understanding of CTSs. This assessment would, however, increase the length of the questionnaire and represent a challenge for future research on how to reliably gather this information.

Second, participating in a choice experiment is an unfamiliar task for most respondents and involves the realisation of many trade-offs, often regarding an unfamiliar good, which may easily induce errors. Here, it is important to consider that respondents are often facing a DCE for the first time and, thus, it is more difficult for them to fully comprehend information about something they have not yet experienced. Therefore, information provided through *ex-ante* instruments may be misleading and ineffective. In this context, *ex-post* mitigation gives respondents a 'second chance' to rate or rethink how they performed in the task, thus allowing them to scrutinise their initial choices and identify situations in which they exceeded their maximum WTP. This is an advantage, because it better positions respondents to notice possible issues that may have arisen while completing the choice tasks. Longo et al. (2015) made similar observations; they noted that when respondents are given the opportunity to revise their WTP answers, sequence effects in contingent valuation are mitigated. They concluded that the revision of answers allowed respondents to express more informed choices. Respondents may regret their initial choices after they have fully understood the contingent market and after they have had some time to think about the questions, to compare all the programmes, and to assess the outcome of their choices with regard to their disposable income. In this sense, *ex-post* mechanisms mimic a supermarket checkout queue where shoppers may think more carefully about the goods in their cart; that is, whether they want to keep all products, change some, or purchase additional ones.

Third, while the *ex-ante* approach only has an indirect relationship to the stated WTP value, *ex-post* instruments may directly affect the stated WTP. In this sense, *ex-ante* approaches can be considered as 'soft' instruments for controlling HB, in contrast to the 'hard' *ex-post* mechanisms. Given that HB is linked to WTP, methods that interact directly with the sources of HB are more likely to influence (reduce) it. The specific *ex-post* mechanism employed in our study allows respondents to revise their choices if WTP is possibly overstated. Several intrinsic or extrinsic motivations may guide their revisions, such as moral commitments, morality and fairness (Hollander-Blumoff, 2011). This is in line with the dissonance minimisation approach that Morrison and Brown (2009) observed, in which respondents overstated their WTP due to the desire to express support for the good or to nurture a generous self-image. It is also in line with Johansson-Stenman and Svedsäter's (2012) theory with respect to the connection between the magnitude of HB and the moral implications of the good.

Ex-post analysis is also an instrument analysts can use to determine the quality of respondents' choices. By obtaining information about respondents' 'understanding and confidence' regarding their choices, the analyst can apply different analytical and methodological tools to correct, weight or even exclude unreliable responses. However, this approach is not free of criticism, given the lack of a theoretical foundation to guide analysts in the selection of the threshold to be used in the calibration function. When measuring respondents' certainty on a 0–10 scale, Champ et al. (1997) and Blumenschein et al. (2001) observed that HB was eliminated when WTP estimation only included those respondents with a score of 10, whereas Norwood (2005) and Poe et al. (2002) found that scores of 8 and 7, respectively, were the preferred values. Thus, the threshold differs between studies, and threshold selection is somewhat arbitrary (Morrison and Brown, 2009). Additionally, the different ways in which certainty can be measured may also have a bearing on outcomes (Beck et al., 2013). In this respect, the approach that this study follows is free from any requirement to calibrate WTP using respondents' certainty. Instead, it relies on respondents' decisions about whether to reduce their WTP if their initial choices were found to be inconsistent.

The *ex-post* procedure used in this study also allows the analyst to shed light on the likely causes of the original choice inconsistencies. Through the *ex-post* mechanism, we were able to differentiate those respondents who unconsciously chose alternatives with higher associated prices relative to their maximum stated WTP (choices assumed to be related to HB) from those who intentionally confirmed inconsistent choices, probably to influence the study's policy outcomes (choices assumed to be related to strategic bias [SB]), as well as from those respondents who made choices that, for any other reason, should not be considered meaningful. We observed that both types of respondents were present in our study, which confirms the existence of both HB and SB. Both biases impact the estimated WTP and are intertwined. However, whilst HB has a significant effect on mean WTP, the impact of SB was significant only if a 'hard' criterion was assumed that considers all respondents who violated their stated maximum WTP in the choice tasks to behave strategically. There are arguments for using a 'soft' cutoff approach, which considers that respondents behave strategically if they violate their stated cutoffs according to a threshold defined by the analyst. First, the good under study was unfamiliar to the respondents, while the mechanisms and efficacy of the policies on the attributes being considered are also unspecified. Thus, a certain degree of uncertainty in the stated WTP can be expected, especially given the short time within and the conditions under which the data are collected. Second, individuals derive utility from supporting public causes and may be willing to violate the price attribute as one of many attributes on the choice card if the remaining attributes in an alternative are evaluated favourably. Third, the payment card format may have disclosed a maximum stated WTP value that is lower than their actual WTP, with the effect of increasing the incidence of inconsistent choices.

Our study is one of the first applications of the LML model and thus makes a contribution to econometric advances in the SP literature. We find that allowing for more flexible model distributions than the standard MXL model does not change the general conclusions. Despite differences in WTP estimates derived from the two model types (which are not statistically significant overall), we note that LML-based estimates are associated with lower estimated standard errors than their MXL counterparts. At the same time, our experience with the new model calls for caution. Despite the relatively quick estimation of a single model run, the model had to be estimated multiple times to investigate its stability and ensure correct specification and convergence. In addition, the LML-based estimation requires several arbitrary assumptions that we have found to influence the results. These limitations and the critical influence of arbitrary decisions in the estimation process deserve future research before the LML model can become a state of practice for discrete choice models.

We acknowledge several limitations of our study. First, the absence of follow-up questions on consequentiality limits our understanding of how respondents perceived the survey. Despite efforts to inform the respondents that the study's results will be used to tailor future agri-environmental policies and that respondents' opinions matter for the development of these policies, we cannot be certain that all individuals understood and/or believed this information. Although the experiment was carried out by the regional government's agrarian research institute, which is known to be the body that advises governments about agricultural taxes and subsidies, we acknowledge that it may not be sufficient to remind respondents about consequentiality and that it is necessary to expend greater effort to test whether and how consequentiality with respect to payment and policy has been perceived (Needham and Hanley, 2020; Zawojska et al., 2019).

Second, we employed elicitation formats that were not incentive compatible. Using more than two alternatives has been observed to provide lower welfare estimates relative to having only one alternative plus the status quo (Weng et al., 2021). Further, the use of payment cards to elicit the price cutoff may have provided a lower estimate of WTP. Vossler and Holloway (2018) found that WTP using a payment card mechanism was significantly lower than the WTP derived from a theory-driven single binary choice experiment. Future research should test the impact of the proposed *ex-post* mitigation strategy in the context of different elicitation formats, comparing the performance under formats that are incentive compatible and under those that are not. This would inform the analyst whether the bias resulting from using statistically more efficient choice designs outweighs the bias resulting from theoretically incentive-incompatible designs. Third, we did not enquire about the reasons respondents decided to retain their initial choices in the revision process. Asking participants about strategic voting (or attitudes correlated with such behaviour) may be very informative for the interpretation of results, thus clarifying the underlying causes of choice inconsistency and helping analysts in the *ex-post* treatment of data.

To conclude, we affirm that HB may significantly affect WTP estimates for unfamiliar goods. *Ex-ante* and *ex-post* mitigation strategies have been of varying effectiveness in the SP literature. Irrespective of the approach used to reduce HB, our findings demonstrate a need to extend the research effort beyond employing *ex-ante* scripts in experimental tests of their effectiveness and gathering *ex-post* information to investigate their potential to assist with WTP adjustments. Specifically, it is necessary to develop a common understanding of *ex-post* instruments across respondents in order to facilitate more rigorous tests of their effectiveness based on theoretical expectations.

ACKNOWLEDGEMENTS

This research was co-financed by the Andalusian Institute of Agricultural Economics and Training (IFAPA) project PP.AVA.AVA2019.031 and by the European Regional Development Fund (ERDF). MC acknowledges the support of the National Science Centre of Poland (Opus, 2017/25/B/HS4/01076). WB acknowledges the support of the Foundation for Polish Science (FNP) and the National Science Centre of Poland (Preludium, 2016/21/N/HS4/02094). KG acknowledges funding through the Scottish Government Rural Affairs and the Environment Portfolio Strategic Research Programme 2016–2021 (Theme 1: Natural Assets).

ORCID

Sergio Colombo https://orcid.org/0000-0001-7307-9811
Wiktor Budziński https://orcid.org/0000-0001-9866-4267
Mikołaj Czajkowski https://orcid.org/0000-0001-5118-2308
Klaus Glenk https://orcid.org/0000-0002-2472-0235

REFERENCES

Aadland, D. & Caplan, A.J. (2003) Willingness to pay for curbside recycling with detection and mitigation of hypothetical bias. *American Journal of Agricultural Economics*, 85, 492–502.

Akter, S. & Bennett, J. (2013) Preference uncertainty in stated preference studies: facts and artefacts. *Applied Economics*, 45, 2107–2115.

Alemu, M.H. & Olsen, S. (2018) 'Linking Consumers' food choice motives to their preferences for insect-based food products: an application of integrated choice and latent variable model in an African context. *Journal of Agricultural Economics*, 70, 241–258.

Ami, D., Aprahamian, F., Chanel, O. & Luchini, S. (2011) A test of cheap talk in different hypothetical contexts: the case of air pollution. *Environmental and Resource Economics*, 50, 111–130.

Ariely, D., Loewenstein, G. & Prelec, D. (2003) Coherent arbitrariness: stable demand curves without stable preferences. *Quarterly Journal of Economics*, 118, 73–105.

Azevedo, C.D., Herriges, J.A. & Kling, C.L. (2003) Combining revealed and stated preferences: consistency tests and their interpretations. *American Journal of Agricultural Economics*, 85, 525–537.

Bansal, P., Daziano, R.A. & Achtnicht, M. (2018a) Comparison of parametric and semiparametric representations of unobserved preference heterogeneity in logit models. *Journal of Choice Modelling*, 27(1), 97–113.

Bansal, P., Daziano, R.A. & Achtnicht, M. (2018b) Extending the logit-mixed logit model for a combination of random and fixed parameters. *Journal of Choice Modelling*, 27(1), 88–96.

Barrage, L. & Lee, M.S. (2010) A penny for your thoughts: inducing truth-telling in stated preference elicitation. *Economics Letters*, 106, 140–142.

Bazzani, C., Palma, M.A. & Nayga, R.M. (2018) On the use of flexible mixing distributions in WTP space: an induced value choice experiment. *Australian Journal of Agricultural and Resource Economics*, 62(2), 185–198.

Beck, M.J., Fifer, S. & Rose, J.M. (2016) Can you ever be certain? Reducing hypothetical bias in stated choice experiments via respondent reported choice certainty. *Transportation Research Part B*, 89, 149–167.

Beck, M.J., Rose, J.M. & Hensher, D.A. (2013) Consistently inconsistent: the role of certainty, acceptability and scale in automobile choice. *Transportation Research Part E*, 56, 81–93.

Bishop, R.C. & Boyle, K.J. (2019) Reliability and validity in nonmarket valuation. *Environmental and Resource Economics*, 72, 559–582.

Blomquist, G.C., Blumenschein, K. & Magnus, J. (2009) Eliciting willingness to pay without bias using follow-up certainty statements: comparisons between probably/definitely and a 10-point certainty scale. *Environmental and Resource Economics*, 43(4), 473–502.

Blumenschein, K., Blomquist, G.C., Johannesson, M., Horn, N. & Freeman, P. (2008) Eliciting willingness to pay without bias: evidence from a field experiment. *Economic Journal*, 118, 114–137.

Blumenschein, K., Johannesson, M., Blomquist, G.C., Liljas et al(1998) Experimental results on expressed certainty and hypothetical bias in contingent valuation. *Southern Economic Journal*, 65(1), 169–177.

Blumenschein, K., Johannesson, M., Yokoyama, K.K., and Freeman et al(2001) Hypothetical versus real willingness to pay in the health care sector: results from a field experiment. *Journal of Health Economics*, 20(3), 441–457.

Bonsall, P. & Lythgoe, B. (2009) Factors affecting the amount of effort expended in responding to questions in behavioral choice experiments. *Journal of Choice Modelling*, 2(2), 216–236.

Bosworth, R. & Taylor, L.O. (2012) Hypothetical bias in choice experiments: is cheap talk effective at eliminating bias on the intensive and extensive margin of choice. *The B. E. Journal of Economic Analysis and Policy*, 12, 1–26.

Broadbent, C.D. (2014) Evaluating mitigation and calibration techniques for hypothetical bias in choice experiments. *Journal of Environmental Planning and Management*, 57(12), 1831–1848.

Brooks, K. & Lusk, J.L. (2010) Stated and revealed preferences for organic and cloned milk: combining choice experiment and scanner data. *American Journal of Agricultural Economics*, 92(4), 1229–1241.

Brown, T.C., Ajzen, I. & Hrubes, D. (2003) Further tests of entreaties to avoid hypothetical bias in referendum contingent valuation. *Journal of Environmental Economics and Management*, 46(2), 353–361.

Bush, G., Colombo, S. & Hanley, N. (2009) Should all choices count? Using the cut-offs approach to edit responses in a choice experiment. *Environmental and Resource Economics*, 44(3), 397–414.

Caputo, V., Scarpa, R., Nayga, R.M. & Ortega, D.L. (2017) Are Preferences for Food Quality Attributes Really Normally Distributed? An Analysis Using Flexible Mixing Distributions. *Journal of Choice Modelling*, 28, 10–27.

Carlsson, F., Frykblom, P. & Lagerkvist, C.J. (2005) Using cheap talk as a test of validity in choice experiments. *Economics Letters*, 89, 147–152.

Carlsson, F., Mørkbak, M. & Olsen, S. (2012) The first time is the hardest: a test of ordering effects in choice experiments. *Journal of Choice Modelling*, 5(2), 19–37.

Carson, R.T. & Groves, T. (2007) Incentive and informational properties of preference questions. *Environmental and Resource Economics*, 37, 181–210.

Champ, P.A., Bishop, R.C., Brown, T.C. & McCollum, D.W. (1997) Using donation mechanisms to value nonuse benefits from public goods. *Journal of Environmental Economics and Management*, 33(2), 151–162.

Champ, P.A., Moore, R. & Bishop, R.C. (2009) A comparison of approaches to mitigate hypothetical bias. *Agricultural and Resource Economics Review*, 38, 166–180.

Cherry, T.L., Frykblom, P. & Shogren, J.F. (2002) Hardnose the dictator. *American Economic Review*, 92(4), 1218–1221.

Chowdhury, S., Meenakshi, J.V., Tomlins, K. & Owori, C. (2011) Are consumers in developing countries willing-to-pay more for micronutrient-dense biofortified foods? Evidence from a field experiment in Uganda. *American Journal of Agricultural Economics*, 93(1), 83–97.

Colombo, S., Glenk, K. & Rocamora-Montiel, B. (2016) Analysis of choice inconsistencies in on-line choice experiments: impact on welfare measures. *European Review of Agricultural Economics*, 43(2), 271–302.

Czajkowski, M. & Budziński, W. (2015) An insight into the numerical simulation bias – a comparison of efficiency and performance of different types of quasi Monte Carlo simulation methods under a wide range of experimental conditions. *In: Environmental Choice Modelling Conference*, (Copenhagen, 2015).

Czajkowski, M., Vossler, C.A., Budziński, W., Wiśniewska, A. & Zawojska, E. (2017) Addressing empirical challenges related to the incentive compatibility of stated preference methods. *Journal of Economic Behavior and Organization*, 142, 47–63.

Doyon, M., Saulais, L., Ruffieux, B. & Bweli, D. (2015) Hypothetical bias for private goods: does cheap talk make a difference? *Theoretical Economics Letters*, 5, 749–756.

Faccioli, M., Czajkowski, M., Glenk, K. & Martin-Ortega, J. 2020. Environmental attitudes and place identity as determinants of preferences for ecosystem services. *Ecological Economics*, 174, (2020) article 106600.

FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L. & Dalgleish, T. (2012b) What we say and what we do: the relationship between real and hypothetical moral choices. *Cognition*, 123(3), 434–441.

Fox, J.A., Shogren, J.F., Hayes, D.J. & Kliebenstein, J.B. (1998) CVM-X: calibrating contingent values with experimental auction markets. *American Journal of Agricultural Economics*, 80, 455–465.

Franceschinis, C., Scarpa, R. & Thiene, M. (2017) *A Monte Carlo Evaluation of the Logit-Mixed Logit under Asymmetry and Multimodality*, Working Papers in Economics 17/23. (University of Waikato, 2017).

Gibbard, A. (1973) Manipulation of voting schemes: a general result. *Econometrica*, 41(4), 587.

Glenk, K., Meyerhoff, J., Akaichi, F. & Martin-Ortega, J. (2019) Revisiting cost vector effects in discrete choice experiments. *Resource and Energy Economics*, 57, 135–155.

Granado-Díaz, R., Gómez-Limón, J.A., Rodríguez-Entrena, M. & Villanueva, A.J. (2020) Spatial analysis of demand for sparsely located ecosystem services using alternative index approaches. *European Review of Agricultural Economics*, 47(2), 752–784.

Grebitus, C., Lusk, J.L. & Nayga Jr, R.M. (2013) Explaining differences in real and hypothetical experimental auctions and choice experiments with personality. *Journal of Economic Psychology*, 36, 11–26.

Groothuis, P.A., Mohr, T.M., Whitehead, J.C. & Cockerill, K. (2017) Endogenous consequentiality in stated preference referendum data: the influence of the randomly assigned tax amount. *Land Economics*, 93, 258–268.

Haghani, M., Bliemer, M.C.J., Rose, J.M., Oppewal, H. & Lancsar, E. (2021) Hypothetical bias in stated choice experiments: part I. Integrative Synthesis of Empirical Evidence and Conceptualisation of External Validity. *Journal of Choice Modelling*, 41, 100309.

Hanley, N. & Czajkowski, M. (2019) The role of stated preference valuation methods in understanding choices and informing policy. *Review of Environmental Economics and Policy*, 13, 248–266.

Hollander-Blumoff, R. (2011) Intrinsic and extrinsic compliance motivations: comment on Feldmand Washington University. *Journal of Law & Policy*, 35, 53–67.

Jacquemet, N., Joule, R.-V., Luchini, S. & Shogren, J.F. (2013) Preference elicitation under oath. *Journal of Environmental Economics and Management*, 65, 110–132.

Johannesson, M., Liljas, B. & Johansson, P.O.(1998) An experimental comparison of dichotomous choice contingent valuation questions and real purchase decisions. *Applied Economics*, 30(5), 643–647.

Johansson-Stenman, O. & Svedsäter, H. (2012) Self-image and valuation of moral goods: stated versus actual willingness to pay. *Journal of Economic Behavior & Organization*, 84(3), 879–891.

Johnston, R.J., Boyle, K.J., Adamowicz, W., Bennett, J., Brouwer, R., Cameron, T.A. et al (2017) Contemporary guidance for stated preference studies. *Journal of the Association of Environmental and Resource Economists*, 4, 319–405.

Kang, M.J., Rangel, A., Camus, M. & Camerer, C.F. (2011) Hypothetical and real choice differentially activate common valuation areas. *The Journal of Neuroscience*, 31, 461–468.

Ladenburg, J. & Olsen, S. (2014) Augmenting short cheap talk scripts with a repeated Opt-Uut reminder in choice experiment surveys. *Resource and Energy Economics*, 37, 39–63.

LaRiviere, J., Czajkowski, M., Hanley, N., Aanesen, M., Falk-Petersen, J. & Tinch, D. (2014) The value of familiarity: effects of knowledge and objective signals on willingness to pay for a public good. *Journal of Environmental Economics and Management*, 68, 376–389.

Liebe, U., Glenk, K., Von Meyer-Hofer, M. & Spiller, A. (2019) A web survey application of real choice experiment. *Journal of Choice Modelling*, 33, article 100150.

List, J.A., Sinha, P. & Taylor, M.H. (2006) Using choice experiments to value non market goods and services: evidence from field experiments. *Advances in Economic Analysis and Policy*, 6, 1–37.

Lloyd-Smith, P., Adamowicz, W.L. & Dupont, D. (2019) Incorporating stated consequentiality questions in stated preference research. *Land Economics*, 95, 293–306.

Longo, A., Hoyos, D. & Markandya, A. (2015) Sequence effects in the valuation of multiple environmental programs using the contingent valuation method. *Land Economics*, 91(1), 20–35.

Loomis, J. (2011) What's to know about hypothetical bias in stated preference valuation studies? *Journal of Economic Surveys*, 25, 363–370.

Loomis, J.B. (2014) Strategies for overcoming hypothetical bias in stated preference surveys. *Journal of Agricultural and Resource Economics*, 39, 34–46.

Lusk, J.L. (2003) Effects of cheap talk on consumer willingness- to-pay for golden rice. *American Journal of Agricultural Economics*, 85(4), 840–856.

Lusk, J.L. & Schroeder, T.C. (2004) Are choice experiments incentive compatible? A test with quality differentiated beef steaks. *American Journal of Agricultural Economics*, 86, 467–482.

McFadden, D. & Train, K. (2000) Mixed MNL models for discrete response. *Journal of Applied Econometrics*, 15(5), 447–470.

Meyerhoff, J. & Glenk, K. (2015) Learning how to choose—effects of instructional choice sets in discrete choice experiments. *Resource and Energy Economics*, 41, 122–142.

Meginnis, K., Burton, M., Chan, R. & Rigby, D. (2021) Strategic bias in discrete choice experiments. *Journal of Environmental Economics and Management*, 109, 102163.

Morrison, M. & Brown, T.C. (2009) Testing the effectiveness of certainty scales, cheap talk, and dissonance-minimization in reducing hypothetical bias in contingent valuation studies. *Environmental and Resource Economics*, 44, 307–326.

Moser, R., Raffaelli, R. & Notaro, S. (2014) Testing hypothetical bias with a real choice experiment using respondents' own money. *European Review of Agricultural Economics*, 41(1), 25–46.

Murphy, J.J., Allen, P.G., Stevens, T.H. & Weatherhead, D. (2005) A meta-analysis of hypothetical bias in stated preference valuation. *Environmental & Resource Economics*, 30(3), 313–325.

Murphy, J.J., Stevens, T.H. & Weatherhead, D. (2005) Is cheap talk effective at eliminating hypothetical bias in a provision point mechanism? *Environmental and Resource Economics*, 30(3), 327–343.

Needham, K. & Hanley, N. (2020) Prior knowledge, familiarity and stated policy consequentiality in contingent valuation. *Journal of Environmental Economics and Policy*, 9, 1–20.

Norwood, F.B. (2005) Can calibration reconcile stated and observed preferences? *Journal of Agricultural and Applied Economics*, 37(1), 237–248.
Penn, J. & Hu, W. (2019) Cheap talk efficacy under potential and actual hypothetical bias: a meta-analysis. *Journal of Environmental Economics and Management*, 96, 22–35.
Poe, G.L., Clarke, J., Rondeau, D. & Schulze, W.D. (2002) Provision point mechanisms and field validity tests of contingent valuation. *Environmental and Resource Economics*, 23(1), 105–131.
Ready, R.C., Champ, P.A. & Lawton, J.L. (2010) Using respondent uncertainty to mitigate hypothetical bias in a stated choice experiment. *Land Economics*, 86(2), 363–381.
Revelt, D. & Train, K. (1998) Mixed logit with repeated choices: households' choices of appliance efficiency level'. *Review of Economics and Statistics*, 80(4), 647–657.
Roe, B., Boyle, K.J. & Teisl, M.F. (1996) Using conjoint analysis to derive estimates of compensating variation. *Journal of Environmental Economics and Management*, 31, 145–159.
Rolfe, J. & Wang, X. (2011) Dealing with scale and scope issues in stated preference experiments. In: Bennett, J. (Ed.) *International Handbook on Non-Market Environmental Valuation*. (Cheltenham: Edward Elgar, pp. 254–273).
Rolfe, J. & Windle, J. (2003) Valuing the protection of aboriginal cultural heritage sites. *The Economic Record*, 79, s85–s95.
Sælensminde, K. (2003) Embedding effect in valuation of non-market-goods. *Transport Policy*, 10(2003), 59–72.
Satterthwaite, M.A. (1975) Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10(2), 187–217.
Silz-Carson, K., Chilton, S.M., Hutchinson, W.G. & Scarpa, R. (2020) Public resource allocation, strategic behavior, and status quo bias in choice experiments. *Public choice*, 185(1–2), 1–19.
Sun, L. & Cornelis van Kooten, G. (2009) Comparing fuzzy and probabilistic approaches to preference uncertainty in non-market valuation. *Environmental and Resource Economics*, 42, 471–489.
Swait, J. (2001) A non-compensatory choice model incorporating attribute cut-offs. *Transportation Research B*, 35, 903–928.
Tonsor, G.T. & Shupp, R.S. (2011) Cheap talk scripts and online choice experiments: looking beyond the mean. *American Journal of Agricultural Economics*, 93(4), 1015–1031.
Train, K.E. (2009) *Discrete Choice Methods with Simulation*, 2nd edition. New York: Cambridge University Press.
Train, K. (2016) Mixed logit with a flexible mixing distribution. *Journal of Choice Modelling*, 19, 40–53.
Train, K.E. & Weeks, M. (2005) Discrete Choice Models in Preference Space and Willingness-to-pay Space, In: Scarpa R. and Alberini A. (Eds.) *Applications of Simulation Methods in Environmental and Resource Economics*. (Dordrecht: Springers, pp. 1–16).
Varela, E., Mahieu, P.-A., Giergiczny, M., Riera, P. & Soliño, M. (2014) Testing the single opt out reminder in choice experiments: an application to fuel break management in Spain. *Journal of Forest Economics*, 20, 212–222.
Vossler, A.C. & Holladay, J.S. (2018) Alternative value elicitation formats in contingent valuation: mechanism design and convergent validity. *Journal of Public Economics*, 165, 133–145.
Vossler, C.A., Doyon, M. & Rondeau, D. (2012) Truth in consequentiality: theory and field evidence on discrete choice experiments. *American Economic Journal Microeconomics*, 4, 145–171.
Weng, W., Morrison, M.D., Boyle, K.J., Boxall, P.C. & Rose, J. (2021) Effects of the number of alternatives in public good discrete choice experiments. *Ecological Economics*, 182, 106904.
Whitehead, J.C. & Cherry, T.L. (2007) Willingness to pay for a green energy program: a comparison of ex-ante and ex-post hypothetical bias mitigation approaches. *Resource and Energy Economics*, 29, 247–261.
Zawojska, E., Bartczak, A. & Czajkowski, M. (2019) Disentangling the effects of policy and payment consequentiality and risk attitudes on stated preferences. *Journal of Environmental Economics and Management*, 93, 63–84.
Zawojska, E. & Czajkowski, M. (2017) Re-examining empirical evidence on stated preferences: Importance of incentive compatibility. *Journal of Environmental Economics and Policy*, 6(4), 374–403.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: 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. *Journal of Agricultural Economics*, 00, 1–29. Available from: https://doi.org/10.1111/1477-9552.12484