Journal of Environmental Economics and Management 109 (2021) 102509

Contents lists available at ScienceDirect

Journal of Environmental Economics and Management

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

Can immersive virtual reality increase respondents' certainty in discrete choice experiments? A comparison with traditional presentation formats

Ilias Mokas [a,*], Sebastien Lizin [a], Tom Brijs [b], Nele Witters [a], Robert Malina [a,c]

[a] Research Group Environmental Economics, Hasselt University, Martelarenlaan 42, 3500, Hasselt, Belgium
[b] Transportation Research Institute (IMOB), Hasselt University, Martelarenlaan 42, 3500, Hasselt, Belgium
[c] Laboratory for Aviation and the Environment, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA, 02139, USA

ARTICLE INFO

Keywords:
Discrete choice experiment
Green infrastructure
Hypothetical bias
Scale parameter
Scale heterogeneity
Virtual reality

ABSTRACT

Stated preference methods such as discrete choice experiments (DCEs) are used to elicit respondents' preferences and willingness to pay (WTP) for environmental goods or services whose value cannot be observed in actual markets. However, DCEs may deliver biased estimates because of respondents' unfamiliarity with the hypothetical scenarios to be valued. There is evidence that visualization techniques can enhance respondents' cognitive ability and improve the evaluation and interpretation of complex information. We leverage recent technological advances to create an immersive virtual reality environment delivered to respondents via a head-mounted display in order to conduct a split split-sample experiment on the value of urban greenery (i.e., trees, bio-retention planters) using three different presentation formats (text only; video; virtual reality). We find that (i) respondent certainty can be increased by employing more immersive visualization techniques such as virtual reality, and that (ii) the presentation format has a significant impact on WTP estimates for different types of urban green and can change respondents' rank order for the urban green options considered in the study.

1. Introduction

Valuation methods in environmental economics are divided into two main core categories: Stated preference (SP) and revealed preference (RP) methods. While the RP method relies on observations of consumers' actual behavior for identifying consumers' preferences and the value of an environmental good, the SP method relies on data that comes from consumers' responses to at least partially hypothetical questions, to elicit the value of changes in environment goods (Kjær, 2005).

A widely used SP method is the Discrete Choice Experiment (DCE). Respondents in a DCE are assumed to act rationally, following a choice strategy that maximizes their utility (Hanemann, 1984). However, SP techniques have long been criticized because, in reality, this assumption is frequently violated. For instance, it has been shown that respondents might be uncertain when making choices (Beck et al., 2016; Dekker et al., 2016; Lundhede and Thorsen, 2009). Because respondent uncertainty can lead to biased value estimates and

* Corresponding author. Hasselt University, Martelarenlaan 42, 3500, Hasselt, Belgium.
E-mail addresses: ilias.mokas@uhasselt.be (I. Mokas), sebastien.lizin@uhasselt.be (S. Lizin), tom.brijs@uhasselt.be (T. Brijs), nele.witters@uhasselt.be (N. Witters), robert.malina@uhasselt.be, rmalina@mit.edu (R. Malina).

https://doi.org/10.1016/j.jeem.2021.102509
Received 5 October 2020; Received in revised form 22 June 2021; Accepted 1 July 2021
Available online 7 July 2021

consequently to potentially wrong conclusions from DCEs (Li and Mattsson, 1995), studies in SP have tried: (i) to explain the underlying reasons for respondent uncertainty (Li and Mattsson, 1995; Wang, 1997), and (ii) to account for respondent uncertainty into modeling (i.e., Hensher et al., 2012; Lundhede and Thorsen, 2009; Welsh and Poe, 1998).

In order to explicitly estimate respondents' uncertainty in DCE, follow-up questions are being used, in which respondents are allowed to express their certainty level about a decision (Li and Mattsson, 1995). Similar to contingent valuation literature, post-decisional choice certainty questions are integrated into surveys (i.e., Hensher et al., 2012; Lundhede and Thorsen, 2009), either in the form of a numeric scale or text statements (Olsen et al., 2011). To date, the integration of stated certainty into modeling remains controversial. For instance, Kosenius (2009) and Lundhede and Thorsen (2009) experimented with different recoding approaches. Overall, they found that none of the recoding approaches had a significant impact on the attribute WTP estimates. On the other hand, Hensher and Rose (2012), following a similar approach to Li and Mattsson (1995), weighted the observations according to the corresponding level of self-reported certainty, they showed an improved model fit.

In order to implicitly estimate respondents' uncertainty in DCE, variations in the scale parameter are assessed. Scale heterogeneity, also known as heteroscedasticity, and its importance on modeling human choices has been studied by many researchers (Scarpa et al., 2003); Scale heterogeneity "*is defined as variation across individual decision-makers in the impact of factors that are not included in the model, relative to the impact of factors that are included*" (Hess and Train 2017). Swait and Louviere (1993) and Breffle and Morey (2000) were some of the first authors who provided insights into scale heterogeneity by integrating the scale parameter in their modeling efforts. These studies outline that the scale parameter reflects the variation of randomness from the perspective of the individual decision-maker, which can thereby be (partly) interpreted as the degree of uncertainty for a choice (Thiene et al. 2015) or the analyst's uncertainty in estimating what determines the preferences of the respondent.

Traditionally, most SP studies have described the alternatives and attributes with a text matrix. Hence SP techniques rely on an individual's imagination to recreate an environmental good, which can lead to higher uncertainty due to poor evaluability and lack of realism (Bateman et al., 2009; Farooq et al. 2018); as such, choice tasks may be equivocal and difficult to understand (Cherchi and Hensher 2015). Past studies have tried to quantify the limitations of an individual to process information and the consequences of these limitations on choice behavior. A number of studies have found that increased task complexity (i) is associated with preference ambiguity (Arentze et al., 2003), and (ii) prevents respondents from selecting their most preferred choice in DCE (Swait and Adamowicz, 2001). Other studies have reported that when individuals' uncertainty vis-à-vis their choices increases, they have less deterministic choices (Beck et al., 2013). Several economics papers relate many of the limitations in SP to choice complexity and uncertainty (Champ et al., 1997; Ready et al. 1995). More specifically, Mazzotta and Opaluch (1995) demonstrated that increased complexity increases the stochastic component of the function used to model the respondents' choices.

Visualization techniques such as photos that enhance respondents' cognitive ability and improve the evaluation and interpretation of complex information could be a key element to improve evaluability and mitigate some of the related bias within DCE (Kanninen, 2007). An early study by MacGregor and Slovic (1986) that employed visual tools showed that the judgmental performance of respondents was correlated to different presentation formats. Lipkus and Hollands (1999) demonstrated that pictorial presentation techniques could improve the awareness of (numerical) risk. Uggeldahl et al. (2016) showed that presenting choice set alternatives in picture format reduced the error variance compared with the text survey. Labao et al. (2008) showed for a contingent valuation study that differences between the photograph specifications (i.e., color versus black and white) had influenced respondents' willingness to pay (WTP), finding a higher WTP for the color format. They claim that this was due to better evaluability of information in the color-format case. In contrast, Hurtubia et al. (2015) conducted a DCE to reveal preferences for urban design. They concluded that while images have several advantages, new biases may be induced, i.e., the temperature of colors was found to affect respondents' choices during the photomontage phase.

More recently, some studies rely on dynamic computer-generated environments (videos). A conjoint analysis study by Dijkstra (2003) demonstrated that while a verbal description of a product might lack realism, the use of visual aids, such as video, both improves the understanding of information by the respondents and increases the reliability of the results. Although Orzechowski et al. (2005) found that the multimedia treatment group had a lower error variance, which they interpreted as less inconsistent, they highlighted that there were no important differences between the verbal and multimedia presentation. A seminal experiment was carried out by Bateman et al. (2009), who performed a choice experiment with the aid of a video to elicit preferences for a land-use case.[1] After splitting the respondents into control and treatment groups, they concluded that the "evaluability" of the choice sets was enhanced in the treatment video group. This view is supported by Matthews et al. (2017), who performed a choice experiment analysis to identify the respondents' priorities for different coastal erosion management strategies using video scenarios that were delivered via an online survey. They find that while the photorealism of the presentation format they used was limited, the choice error was reduced in the video treatment sample. In contrast, another DCE study by Rid et al. (2018) showed that the choice sets presented in 3D still images were more comprehensible compared to computer-generated video while analyzing housing development alternatives.

In Virtual Reality (VR), participants experience immersive scenarios through a head-mounted display that blocks off all visual stimuli from the outside world. Participants enter a VR platform and can navigate in the virtual environments with the use of auxiliary

[1] We note that Bateman et al. (2009) describe their video as a "virtual reality" environment, which is in line with the terminology used in the early 2010s for 3D computer environments that are seen by respondents on a standard computer screen. However, the term Virtual Reality is nowadays predominately used for immersive virtual environments that are experienced by means of head-mounted displays that completely block off visual stimuli from the outside world, whereas nowadays the Bateman approach would best be described as a computer-generated video. This is the terminology that we follow in this manuscript.

devices, for example, a controller glove. Recently, VR has become a research tool for scientists to allow them to both convey complex information and control for visual attributes, with various studies from several fields, i.e., experimental economics, architecture, cognitive psychology, and health care, among others, using VR to study topics such as racial prejudice and discrimination (Vang and Fox, 2014), rehabilitation systems and health outcomes (Rose et al. 2018), and preferences for housing design variations (Orzechowski et al., 2000). It has been shown that VR can better inform researchers due to (i) higher immersion of the participants, and (ii) the researchers' higher control over the experimental design process (Cummings and Bailenson 2016). Bohil et al (2011) argue that researchers can manipulate stimulus inputs so that they maximize respondent's illusion of being immersed. The perception of "presence" caused by the immersive experience could activate the participants' senses, allowing them to reveal unconscious physiological behaviors, which are created by the distinct feeling of "being there". The latter is also demonstrated by Xu et al. (2020), who show that participants experienced a high degree of presence in the VR environment when compared to a real-life benchmark.

Specifically with regard to the use of VR in choice modeling, there is a limited but emerging set of experiments that use VR to convey complex information: Patterson et al. (2017) performed a VR DCE to elicit preferences for neighborhood characteristics related to transportation. They suggested that VR can attract respondents' attention and may be employed in research related to urban planning. Birenboim et al. (2019) conducted a conjoint experiment employing a VR environment, finding a greater sense of presence created by the VR environment compared to still images, supporting the idea that VR results in an improved level of realism compared to traditional presentation formats.

In summary, the existing literature points toward presentation formats impacting on respondents' uncertainty and WTP, and there is some limited evidence on the impact of computer-generated videos experienced via standard computer screens on respondents' uncertainty in discrete choice experiments for environmental goods. However, to the best of our knowledge, no study, to date, has addressed respondent uncertainty in choice experiments for environmental goods when participants are using a head-mounted VR display as a novel immersive presentation format, and has compared this uncertainty to the one associated with other presentation formats such as text and computer-generated video. Moreover, as far as we know, no study in the field of environmental economics has, to date, compared VR-elicited willingness to pay (WTP) with that of other presentation formats such as text and video.

Consequently, the main objectives of our study are: i) to quantify the effect of the presentation format on the mean marginal WTP over all respondents, and ii) to quantify both implicitly and explicitly the effect of the presentation format on respondents' (un)certainty: Implicitly, by estimating a pooled model with fully correlated parameters and by quantifying the effect of each presentation format on the error variance. Explicitly, by estimating the self-reported certainty mean for each of the presentation formats and testing for potential differences between certainty means; and by estimating whether high stated certainty is related to a significantly lower error variance for each of the different subsets.

In other words, we are interested in finding out whether there are significant differences in mean marginal WTP estimates elicited by VR compared to other presentation formats, and whether WTP estimates derived from VR can be regarded as more accurate to those from other formats as respondents are less uncertain about their choices.

To do so, we first test whether the presentation format has an effect on respondents' uncertainty, using a comparative study between three different presentation formats (i.e., text, video, VR) with a split-sample approach. We then use the same experimental data to test for variations of respondents' WTP according to the presentation format.

We find that stated certainty is significantly higher for participants in the VR experiment compared to the text-based experiment, that both multimedia presentation formats (video and VR) have a lower associated error variance, with VR displaying the lowest, and that the WTP estimates are significantly influenced by the presentation tool.

The remainder of this paper is organized as follows: Section 2 introduces the theoretical framework used to quantify respondent uncertainty, and explains both the Virtual Reality environment and the econometric model used. Section 3 presents the collected data and reports the regression results and WTP estimates. Section 4 discusses the effects of attributes and presentation formats on model estimates and WTP. Section 5 concludes.

2. Method

2.1. Discrete choice experiment

Traditionally in DCEs, respondents are asked to make a choice between hypothetical alternatives, characterized by several attributes and levels which are determined during the experimental design. DCEs allow the analyst to observe the trade-offs between the attributes and to estimate the marginal rate of substitution. The attributes and levels that are used in our DCE and the relevant methods for the exploration and comparison of the different survey formats employed are described below.

2.1.1. Attribute selection

In order to select the relevant attributes and levels, we followed three main principles. The attributes were policy-relevant (Jeanloz et al., 2016), and reflected public interests and needs, and they could be developed for both text, video, and the VR survey. The latter principle was an important specification since the aim of this study is to compare different presentation formats. Our study emphasizes the appearance of the urban green; therefore, only those attributes that are directly related to the aesthetic perception were selected for the DCE. After selecting a number of relevant attributes, the questionnaire was developed based on consultations with technical and policy experts. During the first focus group discussion (N = 6), in collaboration with the experts (i.e., architects, economists), we narrowed down the attributes to a set. Throughout the second focus group (N = 15), participants were asked to fill out a draft questionnaire and rank the remaining attributes from best to worst. They also provided comments about potential biases, the chosen payment vehicle, the complexity, and the format of the questionnaire. Moreover, as one of the objectives of this paper is to estimate WTP, we also included monetary cost as an attribute. The experts were asked to give their personal opinion on what their WTP is for different reanimation scenarios that were presented to them. The price levels were determined based on their responses, plausibility, and realism. We did this in order to reduce the risk of strategic behavior (Bateman et al., 2002). Finally, based on the review of the descriptions from the experts and the literature, the developed experiment contained the four most substantial attributes and their corresponding levels (Table 1). In this paper, the "tree canopy and spacing", the "bioretention planters" and "side of the street" are specified as dummy variables, and the "cost" as a continuous variable.

2.1.2. *Experimental design*

The number of attributes and their corresponding levels result in 64 possible combinations of hypothetical management options. An important practical implication is that the use of VR did not allow for a large data sample, as respondents had to come to the laboratory to avoid any location bias. Thus, we limited the number of selected attributes to four and applied a Bayesian efficient design aiming to reduce the required sample size. We followed a two-step design approach utilizing the Ngene software ChoiceMetrics (2017) First, an optimal efficient design with 12 choice sets and with an equal representation of each attribute level was created (Rose and Bliemer 2014). Since we had no prior information on the attribute coefficients and no research had attempted to estimate the values for changes in green infrastructure (GI) management scenarios, a pilot study (N = 120) was conducted, and a Conditional Logit model was estimated using R software (Hess and Palma, 2019). In the next step, in order to optimize a reasonable number of survey questions, a Bayesian D-efficient design was employed using the software Ngene ChoiceMetrics (2017). The developed design takes into account the expected sign of the estimated parameters, which also leads to the elimination of dominant alternatives. The Bayesian design was calculated based on Gauss pseudo-random draws, and the primary reason for selecting that approach over a fractional efficient design is the diminished standard error, which not only allows for smaller sample sizes (Bliemer and Rose, 2011), but also produces estimates that are less sensitive to mis-specification of the priors. The procedure above resulted in 12 choice sets and indicated approximately the minimum sample size (N = 52) that was required for the experiment in order to obtain enough statistical power for the hypothesis testing on the model coefficients ChoiceMetrics (2017). For each choice set, a three-option layout was created where respondents had to choose their most preferred option between two future streetscape alternatives and one status-quo option for those who didn't want any additional street intervention. We included the status-quo so that respondents could choose for no additional environmental interventions in their neighborhood, which also entails no additional costs. The status-quo option is often used in DCE in order to increase the realism of the decision as it allows the respondents to stay with the current design. Excluding the status-quo from the design of hypothetical studies has moreover shown an increase in hypothetical bias (Alemu and Olsen, 2018).

2.1.3. *Questionnaire design*

The survey was synthesized in four parts. In the first part, we included guidelines about the structure of the experiment and "warm up" questions about GI. According to Bateman et al. (2002), "warm up" questions allow the respondents to become familiarized with the policy under examination and to comprehend better the choice sets. Secondly, the respondents were introduced to the valuation scenario and the hypothetical market. We highlighted to them that, in general, participants tend to choose a more expensive alternative than they would do in a real situation. We did as such because cheap talk has been shown to reduce hypothetical bias (Cummings and Taylor 1999). The background context of the choices introduced the residents to the experimental scenarios by asking them to imagine that they live in the neighborhood and that the municipality is planning to reanimate the area. Since the purpose of the study was to compare different presentation formats, the participants were divided into three groups to which the display format was randomly assigned: the control text-only group, the first treatment group (video fly-through), and the second treatment group (VR). The respondents were asked to answer 12 choice sets in total. Each choice set consisted of two hypothetical GI strategies that were reflecting the varying values of the selected attributes and the status quo. In the third section, debriefing questions were asked according to Bateman et al. (2002). Finally, the last and fourth section was related to socio-economic characteristics (i.e., income, gender, and age) (Bateman et al., 2002).

Since we were interested in analyzing the individual stated certainty, an additional question at the end of each choice set asked participants to rate how certain they were of their answer on a scale from 1 (very uncertain) to 10 (very certain). The 12 choice questions were presented in a randomized order to reduce potential order bias. The developed survey design was pre-tested on a group of academics and students, and in response to their feedback we created the final survey.

2.1.4. *3D modeling development*

Based on the experimental design and the resulting text-based matrices, corresponding[2] VR scenarios were created with Unity3D software creating an "average-looking Flemish street". It has been shown that because people become attached to certain places (Low and Altman, 1992), the development of a virtual (imaginary) environment could eliminate bias related to the effect of place attachment on people's preferences. The scenarios were always showing the same street from the first-person point of view, only changing each time the selected attributes and levels. Emphasis was given to designing an immersive environment, including 3D objects[3] such as buildings, trees, wind (animation), and city sounds. We produced 3D renderings for both video display and a sequence of immersive

[2] In order to confirm whether respondents find the text-matrix and visual presentation formats to be corresponding, we pilot-tested this with N = 38. All respondents agreed that the text-matrix and the visual formats were corresponding.

[3] Buildings, type of trees, and sounds are reflecting as close as possible the Flanders region in Belgium.

Table 1
Selected attributes and levels.

Attribute	Level	Interpretation
Tree canopy and spacing	1. Many high canopy trees 2. Few high canopy trees 3. Many low canopy trees 4. Few low canopy trees (base level)	The four levels correspond to the combination of two design properties: the tree canopy, which is the shaded area under the tree created by the leaves and the tree-spacing along the streets.
Bio-retention planters	1. Many planters 2. Few planters (base level)	The two levels correspond to the density (many or few) of planters along the street with evergreen plants and shrubs
Side of the street	1. One (base level) 2. Two	Streets are designed to accommodate the side green (trees, shrubs) either from one or both sides of the street
Price	1. €5 2. €10 3. €20 4. €50	Monthly additional environmental tax payment

environments in VR, where participants could experience the different street scenarios through the head-mounted display Oculus Rift. The VR users could enter in the Oculus platform to navigate in the exported game. The direction of view in VR was controlled by the participant's head position, while the movement was controlled with a wireless joystick, which allowed them to navigate freely (by teleport) in the 25 street designs. In addition, a fast teleport point was added, allowing the respondents to be directly transferred to the end of the street. This specific feature was applied because some of the virtual environments were displayed multiple times, and we wanted to reduce the respondents' possible fatigue in the VR space. For the video treatment group, 25 video sequences were created. The rendering files were exported in order to be reproduced as a.mp4 file via the Qualtrics software, and the length of the videos was 18 s. Buttons such as "start" and "pause" were displayed, allowing the users to view the videos again if needed.

In each VR-based choice set, only the attributes of interest differ, while the rest of the infrastructure, including the surrounding buildings, sounds, and animations remain unchanged. To enhance the realism and develop a user-friendly interface, a part of the choice sets was first presented to a number of users (N = 20) whose feedback assisted us to further improve the virtual experience (see Fig. 1). We improved the choice set transition, the (teleport) navigation in VR to reduce possible nausea effects compared to the initial joystick used for navigation, the animation and realism of trees by randomizing them so that each tree looks unique, the explanatory canvases to assist the user in VR, and the lighting, shadowing, and sky, making them more realistic.

2.1.5. Respondent recruitment and questionnaire distribution

The respondents were recruited using two methods: advertising via local newspapers and television, and through emailing local university staff. A total of 180 respondents participated in the survey (aged 18 years and older). The survey was conducted at Hasselt University (Belgium) over a period of 3 weeks in a properly equipped classroom. As an incentive for participation, each respondent was given food, beverages, and a chance to win a €50 voucher. We used a split-sample approach. Participants were randomly allocated to one of the computer screen displays to complete the text-only, the video, or the VR survey. While exactly the same scenarios were presented to all three groups, the intro-phrasing of each VR choice set question was slightly differentiated, providing additional instructions about how respondents should use the head-mounted display. To ensure that respondents remembered the correct order of the alternatives in each choice set, images that were taken from the corresponding virtual environment were illustrated in the questionnaire. An example of a choice question in the three different formats is shown in Fig. 2. The survey was self-administered, but interviewers were present to assist if required. All three surveys could be completed both in English and in Dutch.

The survey was administered to the participants in groups of 5–12, and the survey ran on identical computers with identical screens. Participants were not allowed to communicate with each other.

2.2. Scale heterogeneity

The scale parameter (λ) is an implicit uncertainty measure and is inversely proportional to the variance of the error term (Bateman et al., 2002; Ben-Akiva and Lerman 1985), meaning that the higher the value of λ, the smaller the idiosyncratic error, and therefore the higher the respondents' certainty. Expecting that the variance of the stochastic component will vary across the respondents, one could model scale by including individual characteristics that can influence the error variance (i.e., income) or by using dummy variables to define subsamples. The scale parameter itself cannot be identified directly. It is compared in relative terms between two or more groups of examination. In our study, one control (the text) and two treatment groups (video and VR).

Various choice models exist that allow for scale or preference heterogeneity. One model, which is widely known for allowing preferences to vary, is the random parameter logit model or mixed logit (MXL) model (Train, 1998). However, MXL models that allow for full correlation among coefficients are less common in the literature.

Fig. 1. Participants give their feedback to improve the VR experience.

2.3. Estimation of MXL with scale parameter

In a standard logit model, the observed preferences β do not vary across individuals *n,* and the effects of unobserved preferences and attributes are captured by ε_{nit}. However, the unobserved preference heterogeneity in discrete choice modeling is an important source of error variability. The MXL model takes into account the respondents' preference heterogeneity as it allows marginal utility coefficients to be distributed randomly between respondents (Hess, 2010). The β' *s* are distributed with density $(\beta|\mathrm{b},)$ and the true utility of each individual *n*, for the alternative *i* in the choice set *t*, is denoted as:

$$U_{nit} = V_{nit} + \varepsilon_{nit} = \beta' x_{nit} + \varepsilon_{nit} \tag{1}$$

where utility (U) is a function of: the error term $\varepsilon_{\mathrm{nit},}$ which is assumed to have an IID extreme value type I distribution, meaning that there is no correlation in the stochastic term across alternatives and across choices; and the deterministic component $V = f(\mathrm{x_{nit}}, \mathrm{y_n}, |\beta)$, where f is a function which includes x_{nit}, a vector of the attribute levels of the alternatives (urban green characteristics), and the alternative-specific constant (ASC) that controls for left-right bias and y_n which includes the participants' socio-economic characteristics. Further interpretations of the model are discussed by Ben-Akiva and Bolduc (1996) and Revelt and Train (1998). The Gumbel distribution is assumed to allow the researcher to measure the utility function, as the true utility function cannot be directly estimated. Introducing the scale parameter λ_n, the utility of an individual *n* can be adjusted as:

$$U_{nit} = \lambda_{nt} V_{nit} + \varepsilon_{nit} \tag{2}$$

The resulting unconditional probability of choosing alternative *i* over another, *j*, for a MXL model is given by:

$$P_{ni} = \int \left\{ \left[\prod_{t=1}^{T} \left(\frac{exp\left(\lambda\left(\beta'_n x_{nit}\right)\right)}{\sum_j exp\left(\lambda\left(\beta'_n x_{njt}\right)\right)} \right) \phi(\beta_n | b, W) \right] d\beta \right\} \tag{3}$$

where $\phi(\beta|b, W)$ reflects a density function with mean b and covariance W (Train, 2009), and the random β' s allow for preference variation across individuals.

The model contains the scale parameter λ which is inversely related to the variance of the error term in the utility function: as the scale parameter goes up, the variance of the error term goes down and vice versa (Ben-Akiva and Lerman 1985; Swait and Louviere 1993). The relationship between variance and the scale parameter is given by:

$$\lambda_{nt} = e^{(\gamma Q_{nt})} = \frac{\pi^2}{6var(\varepsilon_{nt})} \tag{4}$$

where Q_{nt} is a vector of variables linked to the individual *n* and the choice set *t* that are hypothesized to influence the scale parameter; and where γ is a vector for the corresponding parameters. In this paper, *Q* contains the following variables: the stated choice certainty,[4] the choice set order, and the presentation format.

Given that the analyst introduces a distribution for the random parameters, the model parameters are estimated by maximum simulated likelihood estimation. Essentially, the maximum likelihood tries to find those coefficients that deliver the highest probability in obtaining the observed sample (Grafton et al., 2008):

$$LL = \sum_{n=1}^{N} ln \left(\int \left(\prod_{t=1}^{T} \left(\frac{exp\left(\lambda\left(\beta'_n x_{nit}\right)\right)}{\sum_j exp\left(\lambda\left(\beta'_n x_{njt}\right)\right)} \right) \right) \phi(\beta_n | b, W) d\beta \right) \tag{5}$$

[4] Based on Ethier et al. (2000) and Poe et al. (2002) we used a cut-off point of 7 in the 10-point certainty scale. Respondents whose certainty statement was below 7 were recoded as "less certain", and those who chose above or equal to 7 as "certain".

Now imagine that you are a resident of the street which at the moment has **no urban green.** The municipality is willing to improve the greenery design. **If the following options are the** only **available, which scenario would you prefer the most?**

	Scenario A	**Scenario B**	**No management**
Tree spacing & canopy	few low-canopy trees	many high-canopy trees	none
Side-street lower vegetation	many planters	few planters	none
Green-street design	both sides	one side	none
Cost	€20	€20	none

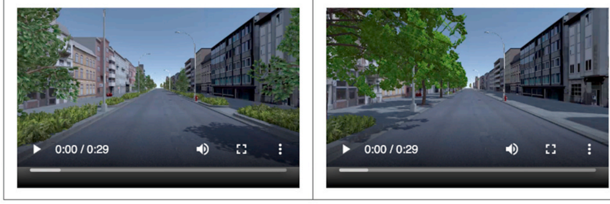
I choose:

Scenario A	**Scenario B**	**No management**
○	○	○

Now imagine that you are a resident of the street which at the moment has **no urban green.** The municipality is willing to improve the greenery design. **If the following options are the** only **available, which scenario would you prefer the most?**

Click on the play button to see the videos below

Scenario A	**Scenario B**	**No management**
€20	€20	None



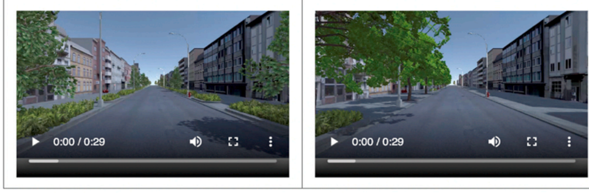
I choose:

Scenario A	**Scenario B**	**No management**
○	○	○

Now imagine that you are a resident of the street which at the moment has **no urban green.** The municipality is willing to improve the greenery design. **The options you will now experience in the Virtual Reality environment are the** only **available. Which scenario would you prefer the most?**

Please put on the VR headset and start exploring the scenarios!

Scenario A	**Scenario B**	**No management**
€20	€20	None



I choose:

Scenario A	**Scenario B**	**No management**
○	○	○

Fig. 2. Example of the choice sets presented to the respondents for the text, video, and VR, respectively.

The simulated probabilities are approximated through simulation for any given value of (b, W).

2.4. WTP space

To derive WTP values for the attributes, a common approach in a DCE is the estimation of models in the preference space using a fixed price coefficient across individuals to avoid complications related to assigning a specific distribution to a random price parameter (i.e., for instance, in a normal distribution, draws close to zero can result in undefined mean WTP for all attributes (Hensher and Greene, 2003; Train and Weeks, 2005). However, fixing the price coefficient is a restrictive assumption as it assumes that the scale parameter is identical for all individuals (Train and Weeks, 2005).

Estimating models in WTP space while allowing for fully correlated random parameters addresses these problems as the mWTP coefficients are directly estimated (Train and Weeks, 2005). Therefore, in these models, it is possible to account both for preference heterogeneity and scale heterogeneity. For this reason, we apply a MXL model in the WTP space with fully correlated random parameters.

More specifically, we specify the model in the WTP-space (Hess and Train 2017):

$$U_{nit} = -\beta_n^p p_{nit} + \beta_n^p wtp_n^{'} X_{nit}^a + \varepsilon_{nit} \tag{6}$$

In the utility expression above, β_n^p is the price coefficient, p_{nit} represents the price attribute, X_{nit}^a is a vector of the non-price parameters, and wtp_n that is defined as $wtp_n = -\beta_n^\alpha / \beta_n^p$ with β_n^p the coefficient of an attribute, reflect the respondents' WTP for the non-price attributes. Using this model specification one can estimate the WTP distribution parameters and interpret them directly.

All utility coefficients, including the ASCs and cost are modelled as random parameters. The non-price parameters were assumed to follow a normal distribution and the price parameter a negative log-normal distribution.

Table 2
Pearson's chi-squared test for socio-economic differences between the control and treatment groups.

	Text		Video		VR		Chi-squared	P-value
	Frequency	Percent	Frequency	Percent	Frequency	Percent		
Age								
Younger than 30	34	56	30	50	28	47		
30–39	15	25	16	27	14	23		
40–49	4	7	6	10	6	10	4.8	0.78(NS)
50–59	2	3	5	8	4	7		
60 yrsand older	5	8	3	5	8	13		
Gender								
Male	28	47	30	50	31	52		
Female	32	53	30	50	29	48	0.31	0.85(NS)
Non-binary	–	–	–	–	–	–		
Education								
Less than high school degree	–	–	1	2	1	2		
High school degree	13	22	8	13	6	10		
Bachelor's degree	11	18	15	25	9	15	7.86	0.44(NS)
Master's degree	25	42	22	37	32	53		
Doctoral degree	11	18	14	23	12	20		
Employment								
Full-time	27	45	35	58	31	52		
Part-time	4	7	4	7	7	12		
Student	22	37	19	32	15	25	11.01	0.3(NS)
Retired	5	8	1	2	6	10		
Self-employed	2	3	–	–	1	2		
Unable to work	–	–	1	2	–	–		
Income								
Less than €1000	5	8	2	3	3	5		
€1000–€1999	3	5	5	8	6	10		
€2000–€2999	16	27	12	20	11	18		
€3000–€3999	6	10	13	22	12	20	19.94	0.06(NS)
€4000–€4999	19	32	16	27	17	28		
€5000–€5999	2	3	9	15	10	17		
€6000 and more	9	15	3	5	1	2		

Note: NS indicates that the probability value is not significant at a conventional 95% level.

3. Results

3.1. Socio-demographics and choice shares

A summary of the socio-economic characteristics of the participants and the distributions of the respondents across the three survey samples is shown in Table 2. In total, 180 participants took part in the experiment with a mean age of 33.71 years (s.d 13.87, range 18–76 years); 49.4% males and 50.6% females. To ensure that the possible differences in the econometric output are not due to different socio-demographic characteristics between the samples, the distributions of each variable for the different splits were tested using the chi-square test. Performing the test, we failed to reject the null hypothesis of equal proportions shown for each socio-demographic variable in each sample (text, video, and VR), which suggests that the three samples are comparable and the differences between groups are associated with the different presentation formats.

3.2. Implicit and explicit quantification of the effect of the treatments on (un)certainty

In order to analyze the 2160 choice observations (180 respondents * 12 choice sets per respondent), nine MXL models were estimated. Table 3 provides an overview of the parameter estimates and model diagnostics. For the initial model results, we pooled the data from all three experiments and we estimated an MXL with all parameters set as random and uncorrelated (Model 1) and two MXL models with random and correlated coefficients (Model 2 and Model 3). Models 2 and 3 incorporate random scale as they allow for

Table 3
Results from modeling scale variation in the MXL pooled models.

	POOLED DATA					
	MXL (Model 1)		MXL with correlated parameters (Model 2)		MXL with correlated parameters, interaction & scale (Model 3)	
	Estimate	t-ratio	Estimate	t-ratio	Estimate	t-ratio
ASC_1	2.6102***	8.39	4.0568***	5.55	3.2968***	4.24
ASC_2	2.7957***	8.98	4.2977***	5.90	3.5733***	4.51
many high canopy	0.8839***	4.69	1.3398***	4.24	1.0271***	3.65
few high canopy	1.2185***	7.13	1.4091***	6.21	1.3387***	5.69
many low canopy	1.1250***	6.45	1.2164***	5.16	1.0710***	4.53
few low canopy (base level)	–	–	–	–	–	–
many planters	0.4611***	4.45	0.7105***	4.31	0.7278***	4.79
few planters (base level)	–	–	–	–	–	–
one side (base level)	–	–	–	–	–	–
both sides	2.0652***	11.40	2.1084***	8.16	1.9475***	6.45
price(ln)	−1.9766***	−24.75	−1.7392***	−14.61	−1.7527***	−13.62
Interactions	–	–	–	–		
ASC_text	–	–	–	–	1	–
ASC_video	–	–	–	–	0.6249	1.04
ASC_VR	–	–	–	–	1.7331**	2.61
S.D						
Sd_ASC_1			0.3777	0.84	1.0023**	2.90
Sd_ASC_2			0.0886	0.50	0.0527	0.23
Sd_many high canopy	1.4291***	7.53	1.9618***	6.89	2.1077***	6.53
Sd_few high canopy	0.8292***	4.24	0.5824*	2.17	0.7302**	3.22
Sd_many low canopy	0.7818***	3.75	0.2719	0.94	0.9308***	4.06
Sd_few low canopy	–	–	–	–	–	–
Sd_many planters	0.6484***	4.68	0.3896*	2.51	0.0244	0.16
Sd_few planters	–	–	–	–	–	–
Sd_one side	–	–	–	–	–	–
Sd_both sides	1.2217***	9.17	0.6117*	2.46	1.2463***	4.32
Sd_price	0.7216***	12.61	0.0056	0.08	0.4612***	5.89
Scale						
Text	–	–	–	–	1	–
Video	–	–	–	–	1.0149***	6.18
VR	–	–	–	–	1.3179***	5.22
Model fit						
Observations	180		180		180	
LL	−1218		−1174		−1160	
Adj-ρ^2	0.48		0.4868		0.491	
AIC	2463.64		2435.55		2415.61	
BIC	2543.13		2685.38		2688.15	

AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion.
Note: * indicates significance at 95% level, ** at 99% level and *** at 99.9% level.

Table 4
T-tests for differences in mean stated certainty between the presentation formats.

	Text vs. Video		Text vs. VR		Video vs. VR	
	D-value	t-value	D-value	t-value	D-value	t-value
Stated certainty	0.12	0.65	0.35*	2.017	0.23	1.304

Note: * indicates significance at 95% level, ** at 99% level and *** at 99.9% level.

correlation between all parameters. The differences between Model 2 and Model 3 are that Model 3 includes additional interaction variables between the presentation format and the ASCs, and that it also allows for scale differences due to the experiment type. All choice models were estimated using the R software (Hess and Palma, 2019) with 1000 Halton draws which were sufficient to stabilize the results.[5]

The estimated magnitude of the coefficients reflects the marginal impact on utility. Moreover, a large standard deviation (Sd) of a random parameter indicates a wide variation in the participants' preferences. For example, in Model 1, we observe a strong preference for "few high canopy trees" for "both sides" of the street; and the random variable "Sd_many high canopy" is both significant and larger than the other random variables, indicating wide variation in preferences for "high canopy". Similar results are found in Model 2 and 3 that allow also for correlation between the parameters.

The ASC were included in the models to capture the average effect of the unobserved factors that are not included (Train, 2009). While a statistically significant and positive parameter estimate for ASC indicates that the respondents favor the "no-management" option, adding two alternative-specific constants indicates whether the participants are more likely to choose the depicted left or right scenario. Both constants are statistically significant (for all models) and the magnitudes of the coefficients are not statistically significantly different, meaning that we do not find evidence respondents favor the left or right alternative. In addition, the interaction term between the ASC and the presentation format was significant only for the interaction between the ASC and the VR indicating that the VR group is more likely to prefer an alternative scenario over the status-quo.

For the pooled models (Models 1, 2 and 3), all estimates for the GI characteristics are statistically significant at the 1% level. Positive signs were obtained for coefficients associated with GI attributes, as people are expected to have a higher preference for a street with more urban green. A negative sign was obtained for "price". Allowing for full correlation amongst coefficients increases the simulated log-likelihood from −1218 (Model 1) to −1174 (Model 2) and to −1160 (Model 3) in the case where we allow for scale to differ between the different presentation formats. Similarly, the adjusted ρ^2 results indicate that Model 2 and Model 3 fit the data better compared to Model 1. Moreover, when estimating the scale parameter of video and VR relative to the text subset (Model 3), we see scale differences (Fig. 3). Scale is statistically significant and has a positive sign, both for video and VR subsets compared to the text sample, implying lower error variance.

Turning to the explicit quantification of respondents' (un)certainty, we first calculated the mean of the stated certainty measure for a particular participant and we used this participant-level measure as an independent unit of observation to create the boxplot for each presentation format (see Fig. 4). This was done in order to account for the fact that the 12 statements of response certainty (i.e., one for each choice set) from a particular respondent are correlated. Visually, the results suggest that the average certainty scores between the control and the two treatment groups are similar, with mean stated certainty being lower in the text setting and higher in the VR setting.

To quantify the effect of presentation formats on choice certainty, we test whether the true difference in mean stated certainty between the different presentation formats is equal to 0, while the alternative hypothesis suggests that the true difference in means is different to 0.

The results for text versus VR suggest that the difference in means is statistically significant and that there is evidence that the mean choice certainty of the VR sample is larger compared to the mean choice certainty of the text sample. On the other hand, t-values of 0.65 and 1.304 are demonstrating that the distributions of choice certainty are similar for the text versus video and video versus VR, respectively (see Table 4).

Additionally, to more precisely capture the effect of respondent certainty and choice set order, we further parameterize choice certainty and choice set order through a scale function for each of the three subsets of data (text, video and VR) (see Table 5). Similar to Models 2 and 3 presented in Table 3, the models displayed in Table 5 are estimated with fully correlated parameters. For all the random coefficients we assume that they are normally distributed except the cost attribute that follows a log-normal distribution. In addition, Model 5, 7 and 9 account for stated certainty and choice order for each of subsets, respectively.

The estimated positive parameters for the stated certainty found in all conditions confirm that choices where respondents are certain result in smaller error variance. When comparing Model 5 (text), Model 7 (video) and Model 9 (VR), we see that for the video treatment when a certainty scale is added, most of the other explanatory variables become statistically insignificant. This implies that certainty has a strong effect on the estimation of the model. Moreover, in all conditions, the estimated negative, but insignificant,

[5] The estimation of the MXL models with correlated random parameters require the estimation of a much larger number of parameters compared to a MXL without correlated parameters. This increases significantly the computational time. Therefore, we tested for differences in model results moving from 500 to 1000 Halton draws. Results showed that the estimated parameters and model diagnostics were relatively stable across the models and, thus, we did not further increase the number of Halton draws beyond 1000.

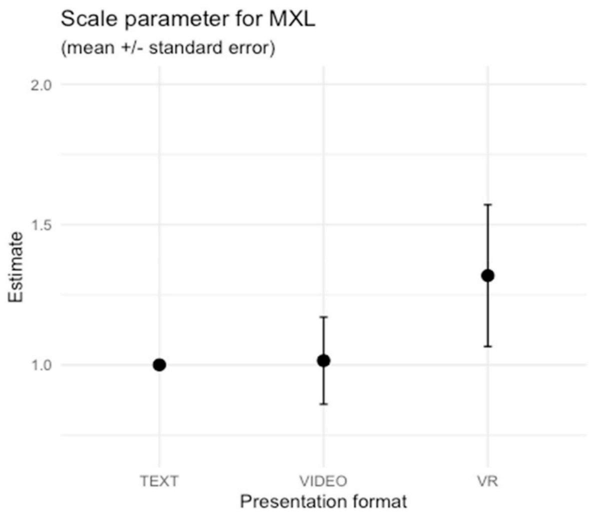


Fig. 3. The scale parameter is increasing from text to video and from video to VR in MXL with correlated parameters.

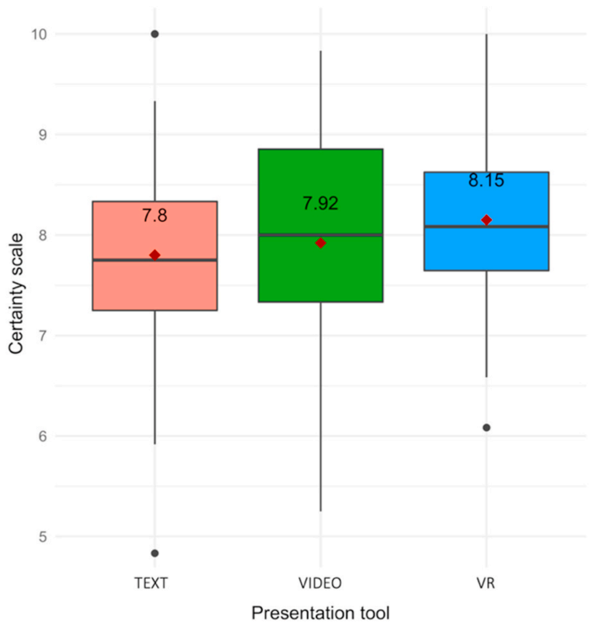


Fig. 4. Boxplot of stated certainty for the different presentation formats.

parameter for choice set order[6] indicates a weak evidence of a possible fatigue effect (Bradley and Daly, 1994).

Looking at the Sd coefficient estimates, for example, for the GI option for "many planters", we see that Sd is reduced from the Model 4 (text) to the Model 8 (VR). This implies that for this GI option, we see a less heterogeneity in preferences in the VR group compared to the text group. On the other hand, the Sd for the GI option "both sides" is larger in the text group compared to the VR group, which implies a larger heterogeneity in preferences for the text group. In addition, comparisons of the distribution of the random correlated price coefficients across the presentation formats show that both for video and VR there is a wider variation in the "price" compared to the text.

We conduct a likelihood ratio (LR) test to test the equality of preferences between samples of data (Swait and Louviere, 1993). Since all the models share common distributional assumptions, we can compare the simulated at convergence log-likelihood (LL) estimates. The LL test is calculated as twice the difference of the pooled log-likelihood value and the sum of the subset log-likelihood values:

[6] The choice set order is coded as an integer from 1 to 12.

Table 5
MXL results for the subgroups.

	Text				Video				VR			
	MXL with correlated parameters (Model 4)		MXL with correlated parameters and scale (Model 5)		MXL with correlated parameters (Model 6)		MXL with correlated parameters and scale (Model 7)		MXL with correlated parameters (Model 8)		MXL with correlated parameters and scale (Model 9)	
	Estimate	t-ratio	Estimate	t-ratio	Estimate	t-ratio	Estimate	t-ratio	Estimate	t-ratio	Estimate	t-ratio
ASC_1	5.0714***	3.89	1.8202***	3.69	3.6332**	2.79	1.3167	1.23	4.6109***	4.16	1.8675*	2.49
ASC_2	5.4475***	4.09	1.9805***	3.91	3.8054**	2.95	1.3546	1.23	4.8642***	4.42	1.7076*	2.46
many high canopy	0.5082	1.18	−0.1438	−1	1.5268**	3.27	0.1685	1.15	4.0050***	7.25	0.6027*	2.49
few high canopy	0.4065	1.27	0.1452	1.42	2.1334***	4.57	0.3582	1.26	3.2475***	6.54	0.7138*	2.53
many low canopy	1.6397***	3.84	0.4739**	3.13	1.3814***	3.40	0.2920	1.21	1.5630*	3.19	0.1305	1.86
few low canopy (base level)	–	–	–	–	–	–	–	–	–	–	–	–
many planters	0.8543**	2.80	0.2065*	2.33	0.6685*	2.41	0.1792	1.28	1.1992***	3.65	0.1638*	2.08
few planters (base level)	–	–	–	–	–	–	–	–	–	–	–	–
one side (base level)	–	–	–	–	–	–	–	–	–	–	–	–
both sides	2.4838***	4.89	1.0423***	4.18	2.2497***	4.98	0.3195	1.24	4.8719***	6.48	1.0424**	2.57
Price(ln)	−1.6413***	−10.05	−2.6993***	−12.71	−1.7174***	−10.56	−3.0034***	−3.83	−1.3170***	−5.87	−2.6011***	−6.80
S.D												
Sd_ASC_1	1.1615*	2.23	0.5019***	4.01	0.2334	0.64	0.0584	0.78	2.2712***	5.56	0.2537*	2.24
Sd_ASC_2	0.1217	0.26	0.0464	0.66	0.1151	0.26	0.0425	0.88	0.4259	1.19	0.0020	0.10
Sd_many high canopy	2.5535***	4.70	0.8807***	3.98	2.1378***	4.36	0.3123	1.23	4.5858***	3.99	1.1011**	2.62
Sd_few high canopy	0.9289*	2.54	0.1947*	2.26	0.6761	1.66	0.2140	1.24	2.0682***	5.15	0.5875*	2.55
Sd_many low canopy	1.3341**	3.41	0.1383	1.11	0.5037	1.24	0.0534	0.98	1.2975***	4.02	0.1739*	2.18
Sd_few low canopy	–	–	–	–	–	–	–	–	–	–	–	–
Sd_many planters	0.7633**	3.00	0.0013	0.02	0.6320*	2.22	0.0249	0.62	0.3333	0.98	0.1861*	2.40
Sd_few planters	–	–	–	–	0.0000	–	–	–	–	–	–	–
Sd_one side	–	–	–	–	0.0000	–	–	–	–	–	–	–
Sd_both sides	0.4001	1.35	0.6143***	3.93	0.6625*	1.99	0.2561	1.28	0.0145	0.04	0.2906*	2.56
Sd_price	0.1673	1.71	0.3081***	5.49	0.5329***	6.22	0.4221***	5.13	0.5451***	5.99	0.4940***	12.41
Scale	–	–	–	–	–	–	–	–	–	–	–	–
Certainty below 7	–	–	1	–	–	–	1	–	–	–	1	–
Certainty over 7	–	–	1.6012***	5.51	–	–	2.6231**	3.32	–	–	3.3994***	6.68
Scale choice order	–	–	−0.9694	−0.79	–	–	−0.0604	−0.20	–	–	−0.0481	−0.39
Model fit												
Observations	60		60		60		60		60		60	
LL	−424		−419		−381		−366		−303		−295	
Adj-ρ^2	0.4086		0.4121		0.4629		0.4792		0.562		0.5686	
AIC	935.6		930.1		849.62		823.86		692.94		682.49	
BIC	1137.09		1140.74		1051.11		1034.5		894.42		893.13	

AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion.
Note: * indicates significance at 95% level, ** at 99% level and *** at 99.9% level.

$$LL_{test} = -2(LL_{A\&B} - (LL_A + LL_B))$$

LL_A and LL_B are the final log-likelihoods for all of the possible combinations of subsets (text, video or VR), and LL_{AB} for the pooled MXL models.[7]

When comparing the log-likelihood values between the pooled models that allow for scale differences, and the presentation format specific models, the hypothesis of equal model fit across samples is rejected at the 0.01% significance level[8] (Table 6). In other words, the estimation of presentation-format specific-models leads to significant gains in model fit when compared to the scaled pooled models. This provides evidence that preferences are different across the three split samples, leading further to the conclusion that the presentation formats do affect the preference coefficients in the choice sets. We note that the beta and scale parameters are confounded, and thus, it cannot be determined whether the differences arise due to the scale parameters, beta parameters or both.

3.3. Marginal willingness-to-pay for the different urban green characteristics

Similar to section 3.2 where we estimated the models in the preference space, allowing for correlation between the parameters, in this section, we estimate the WTP space models. Compared to the preference space models, the WTP space models differ only in the distributional assumptions of the parameters. The preference space models assume normal distribution for all the parameters except for the price that follows a log-normal distribution. On the contrary, in the WTP space, the WTP coefficients are the product of the log-normal with a normal distribution.

Table 7 shows the different mWTP values for the attribute levels that were statistically significant for the three models. For all models, we found a significantly higher mWTP for changing the baseline case into a street with greenery to “both sides”. Having “many bio-retention planters” was the least preferred option when compared with the rest of GI attributes.

Participants who took the text survey are willing to pay more for the “low tree canopy” strategy (10.69 €/month) than the groups who completed the video (6.51 €/month) and VR survey (5.57 €/month). Mean mWTP for the “many high canopy” GI option is significantly larger for the video and VR groups (6.55 €/month and 9.9 €/month respectively) when compared to the text group (3.7 €/month). This indicates a reversal in preferences when going from the text to a multimedia survey. This point is more vivid in Fig. 5 that compares the mean mWTP results (€/month) for the GI attributes across the three experiments. The standard errors are displayed to enable straightforward comparisons among the estimates of the three presentation formats.

A parametric test was conducted to assess whether the mWTP value for each attribute is statistically different among the three groups. Table 8 shows that all the differences - except the ones for “both sides” – are statistically significantly different when comparing the text group to the treatment groups (video and VR). When comparing the two visual groups to each other (video and VR), only one variable was found to be statistically different (“many high canopy”).

4. Discussion

Despite the fact that a standard MXL model with uncorrelated utility coefficients is one of the most frequently applied models in environmental valuation, it is a relatively restrictive model as it implies that the scale parameter is fixed. In this research, a MXL model with fully correlated parameters is specified, allowing for all sources of correlation, including scale heterogeneity.

In our study, we compared the choices of the respondents between the different presentation formats in two ways. First, in an implicit way, by testing whether the scale is significantly affected by the presentation format, and second, in an explicit way, through individual stated choice certainty. The implicit estimation of respondents’ uncertainty provides some evidence for our hypothesis that the visual representation using either video or VR of the choice sets reduces uncertainty. When explicitly estimating the effect of visual representation on respondents’ stated certainty levels, we find that stated certainty between text and VR differs significantly. Moreover, when we capture the effect of respondents’ stated certainty using a scale function approach, our results show that the scale parameter for more certain groups is positive and statistically significant compared to the base (less certain) groups in all conditions. Finally, when comparing the scale coefficient of the group of certain respondents across the three presentation formats, we find that certain respondents in the VR group yield the largest scale parameter.

Taken together, these results provide some support to our hypothesis that VR reduces the randomness in making choices. With regard to the dynamic visualization (video) for the presentation of the choice sets, the present findings are consistent with other research that found that video formats are associated with lower error variance and, therefore, less random responses (Bateman et al., 2009; Matthews et al. 2017; Rid et al., 2018). However, our results suggest that the choice set representation with an immersive VR environment reduced the respondents’ uncertainty, and thus, improved the evaluability for the urban green options compared to a representation with video on a computer screen. Based on our findings, we argue that information completeness regarding the potential strategies can make participants more certain about their choices. A potential explanation for this observation is that while in the text version, the participants have to construct the scenarios in their mind, a representation with video or VR enhances familiarity

[7] To estimate the pooled model we, first, merge the dataset of the subsets A and B (i.e., text and video datasets). Then, a pooled model, based on this merged dataset, which imposes parameter equality is estimated. For the same pooled model, a grid search for the scale parameter that optimizes the log-likelihood is performed.

[8] Since the hypothesis of beta parameter equality is rejected, according to Swait and Louviere (1993), we do not proceed to the second step which is to test for differences in relative scale parameters.

Table 6
LL test shows that preferences are different across the three split samples.

Subset combinations(A)-(B)	LL_A	LL_B	$LL_{A\&B}$ ($\lambda_A \neq \lambda_B$)	LR-test	p-value	Reject H_0: $\beta_A = \beta_B$
(Text) - (Video)	−424	−381	−837	63	0	yes
(Text) - (VR)	−424	−303	−763	72	0	yes
(Video) - (VR)	−381	−303	−702	36	0	yes

Table 7
Average mWTP in €/month across respondents for the different sub-groups.

Variable	Text		Video		VR	
	WTP space Correlation		WTP space Correlation		WTP space Correlation	
	mWTP	t-ratio	mWTP	t-ratio	mWTP	t-ratio
ASC_1	6.12***	4.07	4.48*	2.00	3.39*	2.24
ASC_2	6.1***	4.15	4.59	1.98	3.88*	2.41
many high canopy	3.7**	2.88	6.55***	3.79	9.9***	6.75
few high canopy	2.64**	3.13	7.07***	4.31	7.39***	7.94
many low canopy	10.69***	6.76	6.51***	3.66	5.57***	3.86
many planters	1.38	1.47	3.25***	3.39	3.35***	3.38
both sides	12.7***	12.14	11.45**	2.68	15.87***	7.70
Price(ln)	−1.58***	7.89	1.47***	5.58	0.8**	3.03
S.D						
ASC_1	2.57***	3.69	0.15	0.19	1.75	1.87
ASC_2	0.17	0.66	0.16	0.54	0.36	1.31
many high canopy	11.7***	12.45	9.89***	3.37	13.4***	8.07
few high canopy	1.51	1.91	0.75	0.44	7.07***	6.08
many low canopy	0.51	0.33	3.35**	2.63	4.26**	2.92
many planters	0.67	0.68	3.98**	2.75	3.9***	5.28
both sides	2.505***	4.73	6.85***	3.42	4.11***	5.14
Price(ln)	1.187***	5.87	0.34*	2.50	0.26**	2.59
Model fit						
LL	−435		−387		−314	
Adj-ρ^2	0.3949		0.455		0.5472	
AIC	957.29		862.14		716.37	
BIC	1158.78		1063.63		917.86	

AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion.
Note: * indicates significance at 95% level, ** at 99% level and *** at 99.9% level.

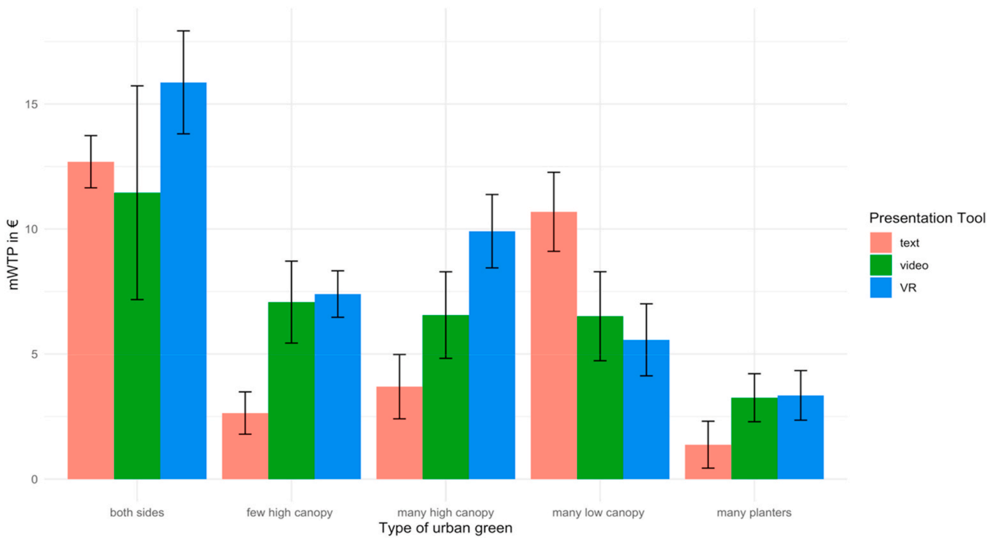


Fig. 5. Mean mWTP by presentation format. Error bars represent the standard errors.

Table 8
T-test to test mWTP differences across experiments.

Variable	Text vs. Video		Text vs. VR		Video vs. VR	
	D-value	t-value	D-value	t-value	D-value	t-value
many high canopy	−2.86***	−2.46	−6.22***	−8.71	−3.36***	−3.656
few high canopy	−4.43**	−3.16	−4.76***	−12.12	−0.32	−0.239
many low canopy	4.18***	5.12	5.12***	7.8	0.94	0.901
many planters	−1.88***	−9.27	−1.97***	−6.17	−0.09	−0.38
both sides	1.24	0.29	3.18	−1.79	−4.42	−1.18

Note: * indicates significance at 95% level, ** at 99% level and *** at 99.9% level.

with the attributes in the choice sets because of the additional visual information provided, which allows to process and evaluate the information more systemically.

Similar to our results, but in a pictorial versus text choice experiment, Uggeldahl et al. (2016) show that pictures in choice set alternatives were associated with lower error variance, arguing that this was due to a decline in complexity of the choice scenarios. Among other studies that have investigated the impact of visual (picture) experiments on the scale parameter, Townsend and Kahn (2014) argue that visual preference heuristics for higher choice complexity cases can induce choice overload compared to textual stimuli, meaning that visual info may also have an adverse effect on randomness. Similarly, Shr et al. (2019) show that the provision of images makes the respondents' answers more random compared to the text sample, and Orzechowski et al. (2005), in a pictorial versus verbal representation of choice sets, state that while some attribute levels may be perceived better with images, they did not reveal any significant differences between the two formats.

The contradictory results in the literature may be explained by the fact that written and visual information is processed differently by different respondents, as some people tend to engage better with a verbal description and others with visual stimuli. This might be an intrinsic, general feature of a respondent, or depend on the object that is valued or both. More specifically, Jansen et al. (2009) state that words may be processed sequentially in a verbal system. Thus, the information could be more rational and logical, whereas images can be handled simultaneously, eliciting in the human brain a direct and almost automatic response of whether or not one likes the image. We argue that, even though much more information is provided in the multimedia representation of the choice sets, some of which may not be relevant, by visualizing a street design, participants are able to understand their preferences better, especially in VR where they can experience the street close to as if it was real. Braga and Starmer (2005) refer to that as "value learning", meaning that simulating goods in an experimental laboratory may help participants to get "closer" to a similar consumption experience.

In addition, we showed that the preference reversal is based on the structure of the decision task, in our case presentation tool, and we attribute this to cognitive and perceptual processing (Howes et al., 2016). It can be argued that the framing and the 3D design for the attributes and levels can influence the results. For instance, people's preferences for greening might be influenced by how nice the high-canopy trees look like compared to the low-canopy. However, as long as the virtual environment is a fair representation of the potential urban green policy, it is prudent to assume that it will more accurately elicit the preferences and WTP of the participants. Respondents are able to experience the potential future policy plans from, literally, different perspectives in a more realistic way.

We would also like to stress that although we used a split sampling approach, and the treatment groups are formed based on participants' shared attributes or characteristics such as income and age, extrapolating the results of our study to the Belgian level should be conducted with great care. Our study is subject to selection bias as people interested in urban green and/or VR were perhaps more likely to participate in the lab-experiment. This could influence the elicited preferences, and consequently the WTP values. Also, highly educated people (above high school diploma) were overrepresented (86% in our sample versus 48.2% in the Belgian population) (Statbel 2018), and are likely to be more concerned about green environments (Gifford and Nilsson, 2014).

Finally, we note that while current VR experiments are limited to laboratory environments, recent technological developments in the field of standalone VR devices and mobile internet will enable researchers to conduct VR experiments outside of the laboratory in the future. This will not only open up the opportunity to mainstream the application of VR in DCE, but will also help to mitigate issues of self-selection bias.

5. Conclusion

In our study, we used a split-sample approach to reveal the effect of presentation formats on respondents' choices using two metrics, which provide implicit and explicit evidence. The implicit approach relied on identifying the scale parameters by means of an MXL that takes into account both preference and scale heterogeneity. The explicit approach relied on self-reported choice certainty that respondents stated after every choice set across all treatments.

To the best of our knowledge, this study is the first attempt to elicit and compare willingness to pay via VR with traditional presentation formats, and to estimate the impact of VR as a novel presentation format on respondents' uncertainty. We highlight three core findings: (1) Stated certainty is significantly higher for participants in the VR experiment compared to the text version. (2) both multimedia presentation formats (video and VR) had a lower associated error variance, with VR displaying the lowest. We conjecture that this could be due to the improved evaluability of the video and VR format that leads to reduced respondent uncertainty when they make choices (3) The mWTP estimates are significantly influenced by the presentation tool.

Focusing specifically on the estimates for the mWTP, the representation of the choice sets in different presentation formats resulted in a reversed rank order for a set of attributes. Such a preference reversal emerged predominately between the text and multimedia (video, VR) representation of the choice set. We explain this with a potential overvaluation or undervaluation of attributes in the less-realistic environments because of respondents' need to rely on their imagination for constructing a visual representation of a scenario.

While we capture preference and scale heterogeneity in our study, quantifying additional forms of heterogeneity, such as processing heterogeneity (see Campbell et al. 2018), could lead to more robust results, reducing at the same time the risk that the unobserved heterogeneity is only partially explained.

Funding source declaration

This work was supported by the Research Foundation Flanders (FWO) [grant numbers 12G5418N, 12B2917N].

Submission declaration and verification

This article was not published previously and it is not being considered anywhere else. All authors approve of the publication.

Declaration of competing interest

Nothing to be declared.

References

Alemu, M.H., Olsen, S.B., 2018. Can a Repeated Opt-Out Reminder mitigate hypothetical bias in discrete choice experiments? An application to consumer valuation of novel food products. Eur. Rev. Agric. Econ. 45 (5), 749–782. https://doi.org/10.1093/erae/jby009.

Arentze, T., Borgers, A., Timmermans, H., DelMistro, R., 2003. Transport stated choice responses: effects of task complexity, presentation format and literacy. Transport. Res. E Logist. Transport. Rev. 39 (3), 229–244. https://doi.org/10.1016/S1366-5545(02)00047-9.

Bateman, I.J., Carson, R.T., Day, B., Hanemann, M., Hanley, N., Hett, T., Jones-Lee, M., Loomes, G., Mourato, S., Özdemirog¯lu, E., Pearce, D.W., Sugden, R., Swanson, J., 2002. Economic Valuation with Stated Preference Techniques: a Manual. Edward Elgar Publishing Ltd.

Bateman, Ian J., Day, B.H., Jones, A.P., Jude, S., 2009. Reducing gain–loss asymmetry: a virtual reality choice experiment valuing land use change. J. Environ. Econ. Manag. 58 (1), 106–118. https://doi.org/10.1016/j.jeem.2008.05.003.

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. Transp. Res. Part B Methodol. 89, 149–167. https://doi.org/10.1016/j.trb.2016.04.004.

Beck, M.J., Rose, J.M., Hensher, D.A., 2013. Consistently inconsistent: the role of certainty, acceptability and scale in choice. Transport. Res. E Logist. Transport. Rev. 56, 81–93.

Ben-Akiva, M., Bolduc, D., 1996. Multinomial Probit with a Logit Kernel and a General Parametric Specification of the Covariance Structure. D'epartement d''economique, Universit'e laval with Department of Civil and Environmental Engineering, Massachusetts Institute of Technology.

Ben-Akiva, M.E., Lerman, S.R., 1985. Discrete Choice Analysis: Theory and Application to Travel Demand. MIT Press. https://books.google.be/books?id=oLC6ZYPs9UoC.

Birenboim, A., Dijst, M., Ettema, D., de Kruijf, J., de Leeuw, G., Dogterom, N., 2019. The utilization of immersive virtual environments for the investigation of environmental preferences. Landsc. Urban Plann. 189, 129–138. https://doi.org/10.1016/j.landurbplan.2019.04.011.

Bliemer, M.C.J., Rose, J.M., 2011. Experimental design influences on stated choice outputs: an empirical study in air travel choice. Transport. Res. Pol. Pract. 45 (1), 63–79.

Bohil, C.J., Alicea, B., Biocca, F.A., 2011. Virtual reality in neuroscience research and therapy. Nat. Rev. Neurosci. 12 (12), 752.

Bradley, M., Daly, A., 1994. Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data. Transportation 21 (2), 167–184. https://doi.org/10.1007/BF01098791.

Braga, J., Starmer, C., 2005. Preference anomalies, preference elicitation and the discovered preference hypothesis. Environ. Resour. Econ. 32 (1), 55–89. https://doi.org/10.1007/s10640-005-6028-0.

Breffle, W.S., Morey, E.R., 2000. Investigating preference heterogeneity in a repeated discrete-choice recreation demand model of Atlantic salmon fishing. Mar. Resour. Econ. 15 (1), 1–20.

Campbell, D., Mørkbak, M.R., Olsen, S.B., 2018. The link between response time and preference, variance and processing heterogeneity in stated choice experiments. J. Environ. Econ. Manag. 88, 18–34. https://doi.org/10.1016/j.jeem.2017.10.003.

Champ, P.A., Bishop, R.C., Brown, T.C., McCollum, D.W., 1997. Using donation mechanisms to value nonuse benefits from public goods. J. Environ. Econ. Manag. 33 (2), 151–162. https://doi.org/10.1006/jeem.1997.0988.

Cherchi, E., Hensher, D.A., 2015. Workshop synthesis: stated preference surveys and experimental design, an audit of the journey so far and future research perspectives. Transportation Research Procedia 11, 154–164. https://doi.org/10.1016/j.trpro.2015.12.013.

ChoiceMetrics, 2017. Ngene 1.2.1, Australia.

Cummings, J.J., Bailenson, J.N., 2016. How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. Media Psychol. 19 (2), 272–309. https://doi.org/10.1080/15213269.2015.1015740.

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

Dekker, T., Hess, S., Brouwer, R., Hofkes, M., 2016. Decision uncertainty in multi-attribute stated preference studies. Resour. Energy Econ. 43, 57–73. https://doi.org/10.1016/j.reseneeco.2015.11.002.

Ethier, R.G., Poe, G.L., Schulze, W.D., Clark, J., 2000. A comparison of hypothetical phone and mail contingent valuation responses for green-pricing electricity programs. Land Econ. 54–67.

Farooq, B., Cherchi, E., Sobhani, A., 2018. Virtual immersive reality for stated preference travel behavior experiments: a case study of autonomous vehicles on urban roads. Transport. Res. Rec. 2672 (50), 35–45. https://doi.org/10.1177/0361198118776810.

Gifford, Robert, Nilsson, Andreas, 2014. Personal and social factors that influence pro-environmental concern and behaviour: A review. Int. J. Psychol. 49 (3), 141–157.

Grafton, Q., Adamowicz, W., Dupont, D., Nelson, H., Hill, R.J., Renzetti, S., 2008. The Economics of the Environment and Natural Resources. John Wiley & Sons.

Hanemann, W.M., 1984. Welfare evaluations in contingent valuation experiments with discrete responses. Am. J. Agric. Econ. 66 (3), 332–341. https://doi.org/10.2307/1240800.

Hensher, D.A., Greene, W.H., 2003. The Mixed Logit model: the state of practice. Transportation 30 (2), 133–176. https://doi.org/10.1023/A:1022558715350.

Hensher, D.A., Rose, J.M., 2012. The influence of alternative acceptability, attribute thresholds and choice response certainty on automobile purchase preferences. J. Transport Econ. Pol. 46 (3), 451–468. http://www.jstor.org/stable/24396322.

Hensher, D.A., Rose, J.M., Beck, M.J., 2012. Are there specific design elements of choice experiments and types of people that influence choice response certainty? Journal of Choice Modelling 5 (1), 77–97. https://doi.org/10.1016/S1755-5345(13)70049-6.
Hess, S., Palma, D., 2019. Apollo: a flexible, powerful and customisable freeware package for choice model estimation and application. Journal of Choice Modelling 100170.
Hess, S., 2010. Conditional parameter estimates from Mixed Logit models: distributional assumptions and a free software tool. Journal of Choice Modelling 3 (2), 134–152. https://doi.org/10.1016/S1755-5345(13)70039-3.
Hess, S., Train, K., 2017. Correlation and scale in mixed logit models. Journal of Choice Modelling 23, 1–8. https://doi.org/10.1016/j.jocm.2017.03.001.
Howes, A., Warren, P.A., Farmer, G., El-Deredy, W., Lewis, R.L., 2016. Why contextual preference reversals maximize expected value. Psychol. Rev. 123 (4), 368.
Hurtubia, R., Guevara, A., Donoso, P., 2015. Using images to measure qualitative attributes of public spaces through SP surveys. Transportation Research Procedia 11, 460–474. https://doi.org/10.1016/j.trpro.2015.12.038.
Jansen, S., Boumeester, H., Coolen, H., Goetgeluk, R., Molin, E., 2009. The impact of including images in a conjoint measurement task: evidence from two small-scale studies. J. Hous. Built Environ. 24 (3), 271–297. https://doi.org/10.1007/s10901-009-9149-x.
Jeanloz, S., Lizin, S., Beenaerts, N., Brouwer, R., Van Passel, S., Witters, N., 2016. Towards a more structured selection process for attributes and levels in choice experiments: a study in a Belgian protected area. Ecosystem Services 18, 45–57.
Kanninen, B.J., 2007. Valuing Environmental Amenities Using Stated Choice Studies: a Common Sense Approach to Theory and Practice, vol. 8. Springer.
Kjær, T., 2005. A review of the discrete choice experiment - with emphasis on its application in health care. Health Economics Papers 1–139.
Kosenius, A.-K., 2009. Causes of Response Uncertainty and its Implica& Tions for WTP Estimation in Choice Experiments. Discussion Papers, vol. 29.
Labao, R., Francisco, H., Harder, D., Santos, F.I., 2008. Do colored photographs affect willingness to pay responses for endangered species conservation? Environ. Resour. Econ. 40 (2), 251–264. https://doi.org/10.1007/s10640-007-9151-2.
Level of education | Statbel (n.d.) Retrieved February 6, 2020, from. https://statbel.fgov.be/en/themes/work-training/training-and-education/level-education#panel-12.
Li, C.-Z., Mattsson, L., 1995. Discrete choice under preference uncertainty: an improved structural model for contingent valuation. J. Environ. Econ. Manag. 28 (2), 256–269. https://doi.org/10.1006/jeem.1995.1017.
Lipkus, I.M., Hollands, J.G., 1999. The visual communication of risk. JNCI Monographs 1999 (25), 149–163. https://doi.org/10.1093/oxfordjournals.jncimonographs.a024191.
Low, S.M., Altman, I., 1992. In: Altman, I., Low, S.M. (Eds.), Place Attachment BT - Place Attachment, 1–12. Springer US. https://doi.org/10.1007/978-1-4684-8753-4_1.
Lundhede, T.H., Thorsen, B.J., 2009. Handling respondent uncertainty in Choice Experiments : evaluating recoding approaches against explicit modelling of uncertainty. Journal of Choice Modeling 2 (2), 118–147. https://doi.org/10.1016/S1755-5345(13)70007-1.
MacGregor, D., Slovic, P., 1986. Graphic representation of judgmental information. Hum. Comput. Interact. 2 (3), 179–200. https://doi.org/10.1207/s15327051hci0203_1.
Matthews, Y., Scarpa, R., Marsh, D., 2017. Using virtual environments to improve the realism of choice experiments: a case study about coastal erosion management. J. Environ. Econ. Manag. 81, 193–208. https://doi.org/10.1016/j.jeem.2016.08.001.
Mazzotta, M.J., Opaluch, J.J., 1995. Decision making when choices are complex: a test of Heiner's hypothesis. Land Econ. 500–515.
Olsen, S.B., Lundhede, T.H., Jacobsen, J.B., Thorsen, B.J., 2011. Tough and easy choices: testing the influence of utility difference on stated certainty-in-choice in choice experiments. Environ. Resour. Econ. 49 (4), 491–510.
Orzechowski, M.A., Arentze, T.A., Borgers, A.W.J., Timmermans, H.J.P., 2005. Alternate methods of conjoint analysis for estimating housing preference functions: effects of presentation style. J. Hous. Built Environ. 20 (4), 349–362.
Orzechowski, Maciej A., Timmermans, H.J.P., De Vries, B., 2000. Measuring user satisfaction for design variations through virtual reality. In: Design and Decision Support Systems in Architecture–Proceedings of the 5th International Conference. Eindhoven University of Technology, The Netherlands, pp. 278–288.
Patterson, Z., Darbani, J.M., Rezaei, A., Zacharias, J., Yazdizadeh, A., 2017. Comparing text-only and virtual reality discrete choice experiments of neighbourhood choice. Landsc. Urban Plann. 157, 63–74. https://doi.org/10.1016/j.landurbplan.2016.05.024.
Poe, G.L., Clark, J.E., Rondeau, D., Schulze, W.D., 2002. Provision point mechanisms and field validity tests of contingent valuation. Environ. Resour. Econ. 23 (1), 105–131.
Ready, R.C., Whitehead, J.C., Blomquist, G.C., 1995. Contingent valuation when respondents are ambivalent. J. Environ. Econ. Manag. 29 (2), 181–196. https://doi.org/10.1006/jeem.1995.1040.
Revelt, D., Train, K., 1998. Mixed logit with repeated choices: households' choices of appliance efficiency level. Rev. Econ. Stat. 80 (4), 647–657.
Rid, W., Haider, W., Ryffel, A., Beardmore, A.B., 2018. Visualisations in choice experiments: comparing 3D film-sequences and still-images to analyse housing development alternatives. Ecol. Econ. 146 (December 2017), 203–217. https://doi.org/10.1016/j.ecolecon.2017.10.019.
Rose, J., Bliemer, M.C.J., 2014. Stated Choice Experimental Design Theory: the Who, the what and the Why. Edward Elgar Publishing.
Rose, T., Nam, C.S., Chen, K.B., 2018. Immersion of virtual reality for rehabilitation - Review. Appl. Ergon. 69, 153–161. https://doi.org/10.1016/j.apergo.2018.01.009.
Scarpa, R., Ruto, E.S.K., Kristjanson, P., Radeny, M., Drucker, A.G., Rege, J.E.O., 2003. Valuing indigenous cattle breeds in Kenya: an empirical comparison of stated and revealed preference value estimates. Ecol. Econ. 45 (3), 409–426.
Shr, Y.-H. (Jimmy), Ready, R., Orland, B., Echols, S., 2019. How do visual representations influence survey responses? Evidence from a choice experiment on landscape attributes of green infrastructure. Ecol. Econ. 156, 375–386. https://doi.org/10.1016/j.ecolecon.2018.10.015.
Swait, J., Adamowicz, W., 2001. Choice environment, market complexity, and consumer behavior: a theoretical and empirical approach for incorporating decision complexity into models of consumer choice. Organ. Behav. Hum. Decis. Process. 86 (2), 141–167.
Swait, J., Louviere, J., 1993. The role of the scale parameter in the estimation and comparison of multinomial logit models. J. Market. Res. 30 (3), 305–314. https://doi.org/10.2307/3172883.
Thiene, M., Scarpa, R., Louviere, J.J., 2015. Addressing preference heterogeneity, multiple scales and attribute attendance with a correlated finite mixing model of tap water choice. Environ. Resour. Econ. 62 (3), 637–656.
Townsend, C., Kahn, B.E., 2014. The "visual preference heuristic": the influence of visual versus verbal depiction on assortment processing, perceived variety, and choice overload. J. Consum. Res. 40 (5), 993–1015. https://doi.org/10.1086/673521.
Train, K.E., 1998. Recreation demand models with taste differences over people. Land Econ. 230–239.
Train, K.E., 2009. Discrete Choice Methods with Simulation. Cambridge university press.
Train, K., Weeks, M., 2005. Discrete choice models in preference space and willingness-to-pay space. In: Applications of Simulation Methods in Environmental and Resource Economics. Springer, pp. 1–16.
Uggeldahl, K., Jacobsen, C., Lundhede, T.H., Olsen, S.B., 2016. Choice certainty in Discrete Choice Experiments: will eye tracking provide useful measures? Journal of Choice Modelling 20, 35–48. https://doi.org/10.1016/j.jocm.2016.09.002.
Vang, M.H., Fox, J., 2014. Race in virtual environments: competitive versus cooperative games with black or white avatars. Cyberpsychol., Behav. Soc. Netw. 17 (4), 235–240. https://doi.org/10.1089/cyber.2013.0289.
Wang, H., 1997. Treatment of "don't know" responses in contingent valuation surveys: a random valuation model. J. Environ. Econ. Manag. 32 (2), 219–232. https://doi.org/10.1006/jeem.1996.0965.
Welsh, M.P., Poe, G.L., 1998. Elicitation effects in contingent valuation: comparisons to a multiple bounded discrete choice approach. J. Environ. Econ. Manag. 36 (2), 170–185. https://doi.org/10.1006/jeem.1998.1043.
Xu, C., Demir-Kaymaz, Y., Hartmann, C., Menozzi, M., Siegrist, M., 2020. The Comparability of Consumers' Behavior in Virtual Reality and Real Life: A Validation Study of Virtual Reality Based on a Ranking Task. Food Quality and Preference, p. 104071.