Estimating decision rule differences between ‘best’ and ‘worst’ choices in a sequential best worst discrete choice experiment

Geržinič, Nejc; van Cranenburgh, Sander; Cats, Oded; Lancsar, Emily; Chorus, Caspar

DOI
10.1016/j.jocm.2021.100307

Publication date
2021

Document Version
Final published version

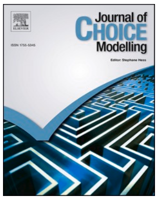
Published in
Journal of Choice Modelling

Journal of Choice Modelling 41 (2021) 100307

Contents lists available at ScienceDirect

Journal of Choice Modelling

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



Estimating decision rule differences between ‘best’ and ‘worst’ choices in a sequential best worst discrete choice experiment

Nejc Geržinič [a,*], Sander van Cranenburgh [b], Oded Cats [a], Emily Lancsar [c], Caspar Chorus [b]

[a] Department of Transport and Planning, Delft University of Technology, Netherlands
[b] Department of Engineering Systems and Services, Delft University of Technology, Netherlands
[c] Department of Health Services Research and Policy, Research School of Population Health, Australian National University, Australia

ARTICLE INFO

Keywords:
Discrete choice model
Best worst discrete choice experiments
Random utility maximisation
Random regret minimisation
Decision rule

ABSTRACT

Since the introduction of Discrete Choice Analysis, countless efforts have been made to enhance the efficiency of data collection through choice experiments and to improve the behavioural realism of choice models. One example development in data collection are best-worst discrete choice experiments (BWDCE), which have the benefit of obtaining a larger number of observations per respondent, allowing for reliably estimating choice models even with smaller samples. In SWDCE, respondents are asked to alternatingly select the ‘best’/‘worst’ alternatives, until the choice set is exhausted. The use of BWDCE raises the question of decision-rule consistency through the stages of the experiment. We challenge the notion that the same fully compensatory decision rule is utilised throughout the experiment. We hypothesize that respondents may utilise one decision rule for selecting the ‘best’ and another for selecting the ‘worst’ alternatives. To test our hypothesis, we developed a model that combines the SBWMNL model for modelling best-worst data and the μRRM model that can account for variations in decision rules. Our results show that decision-rule heterogeneity does seem to be present in BWDCE: it is more likely that ‘best’ choices are made using a fully compensatory decision rule (maximising utility), whereas ‘worst’ choices are more likely made using a non-compensatory decision rule (minimising regret). Such behaviour is largely similar to how image theory describes the decision-making process in complex situations. Our findings give choice modellers new insight into the behaviour of respondents in best-worst experiments and allows them to represent their behaviour more accurately.

1. Introduction

Since the introduction of Discrete Choice Analysis (DCA), it has seen numerous extensions to the data collection process, survey design and modelling approaches, with the goal of developing a more realistic representation of users’ behaviour and to collect data more efficiently. One recent advancement of DCA which has gained attention especially in the field of healthcare are Best Worst Discrete Choice Experiments (BWDCE) (Lancsar et al., 2013). Best worst scaling has been used in research for years (Louviere et al., 2008), but due to its ability to obtain more preference information from respondents, it has become particularly popular in the field of

[*] Corresponding author.
E-mail address: n.gerzinic@tudelft.nl (N. Geržinič).

https://doi.org/10.1016/j.jocm.2021.100307
Received 24 August 2020; Received in revised form 13 April 2021; Accepted 29 July 2021
Available online 7 August 2021

healthcare. As the research population in health-related studies can be small, using traditional DCA would not yield a sufficient number of observations to enable the robust estimation of choice models.

BWDCE are a type of best-worst scaling (Case 3 of best-worst scaling (Flynn, 2010)) that is most similar to traditional Stated Choice (SC) experiments as it includes a set of alternatives, described by their respective attributes; respondents are asked to select the alternative they see as best/worst. BWDCE evolved from ranking experiments – where respondents are simply shown the alternatives and asked to rank them based on their preference – in order to reduce the mental burden of respondents and increase the efficiency of the data collection process (in the sense that from each choice set, multiple observations are obtained), by (1) asking them to select the best/worst alternative in a choice set and (2) removing said alternative. This process iterates until the choice set is exhausted and an implied ranking of the alternatives is obtained. Alternatingly asking respondents to choose the best and worst alternative, rather than continually asking for the best, is also asserted to be easier for respondents as people are believed to identify extremes with less cognitive effort (Marley and Louviere, 2005). Using a data collection technique where respondents are asked to alternatingly select best and worst options, raises the question of whether both decisions are made in the same way: is the decision-making process identical for selecting best and worst alternatives and if not, what kind of implications does that have on the model outcome?

Image theory states that when faced with a decision, people use a two-step decision-making approach to ease the decision-making process. First, decision-makers narrow down the choice set to a more manageable number of alternatives, by removing all options performing below a certain subjective threshold. This is also called '*Compatibility*' and is a non-compensatory decision-making process. Second, decision-makers evaluate the remaining alternatives and trade-off their performances to make a final decision on which alternative to choose. This second step of image theory is also called '*Profitability*' and it is a (fully) compensatory decision-making process, as it involves trading-off of the individual alternatives' performances (Beach and Mitchell, 1987; Meloy and Russo, 2004). Although this is not the exact same order of decision-making as can be observed in BWDCEs, both approaches see decision-makers reduce the size of the choice set through removing best and worst performing alternatives. Both approaches also state a reduced cognitive burden on respondents in this process. We therefore hypothesize that in BWDCEs, like in image theory, different decision rules are used when selecting the best alternative within a choice set, as opposed to selecting the worst alternative.

Recently, a model that can capture the variation in the level of compensatory behaviour - the μRRM model - was proposed by Van Cranenburgh, Guevara and Chorus (2015). It is an evolution of the work on random regret minimisation (RRM) models by Chorus et al. (2008) and G. Chorus (2010). Compared to linear in parameters Random Utility Maximisation (RUM) models, the main distinction of RRM models is that they look at the differences between the alternatives' performance on different attributes. With an additional regret parameter "μ", the μRRM model is able to capture decision rules varying from fully compensatory (as in RUM models) to non-compensatory. Fully compensatory means that a well performing attribute of an alternative can compensate for poor performance of another attribute. For example, a very expensive alternative can still prove attractive if the travel time is short enough. In other words, respondents trade-off the performance of alternatives. Non-compensatory behaviour (referred to as P-RRM or pure RRM) on the other hand means that poor performance of an attribute cannot be compensated by good performance of another, i.e. no matter how short the travel time of an alternative is, it is too expensive compared to other alternatives.

This paper aims to determine if different decision rules are indeed used when choosing the best and worst alternative as stipulated by image theory. We apply a novel modelling approach, where best-worst data are analysed by means of a μRRM model, accounting not only for different levels of utility per stage, but also for different underlying decision rules/different levels of compensatory behaviour. As a modelling approach of this type is new, some model-specific characteristics are explored as well to better understand the behaviour of the model and its outcomes. We test two different approaches of accounting for choice set size variation, as the level of regret in RRM models depends on the number of alternatives present in the choice set. Since the developed models have two parameters varying through the different stages of the choice experiment (the μ regret parameter and the Λ scale parameter accounting for the variation in choice set size), their relationship and impact on the model outcomes are explored as well, and corrected for in the statistical assessment of model fit differences across models.

This paper is structured as follows: Section 2 defines the models that are applied in this research, the statistical tests used to compare the model outcomes and the data collection process. Following in Section 3 are the estimation results of the different models. Finally, in Section 4, the results and limitations of this study are discussed and avenues for follow-up research are presented.

2. Methodology

This section provides a detailed overview of the estimated models and the data collection process. Firstly, the modelling of data collected through BWDCEs is explained in Section 2.1. Image theory and the links between it and the decision-making behaviour in BWDCE is expanded upon in Section 2.2. Section 2.3 then gives a detailed insight into how the variation in the use of decision rules through the BWDCE can be captured by utilising the generalised μRRM model. An overview of the estimated models on the collected data is presented in Section 2.4. Statistical tests used to compare and assess the performance of these different models are explained in Section 2.5. Finally, Section 2.6 provides information on the survey design and the data collection process.

2.1. Analysing best-worst data

The core idea of data collection with the help of BWDCE is obtaining an (implied) ranking of alternatives in each presented choice set. To model this implied ranking, two different approaches are used by researchers as summarised by Lancsar et al. (2013). Both approaches model the probability of the observed ranking in the choice set as a product of individual linear in parameter RUM MNL models (if J is the number of alternatives in the choice set, there are J-1 choices to be made). In a hypothetical set of five alternatives

(ABCDE), where the final ranking is A > B > C > D > E, a Rank Order Logit (ROL) or Exploded Logit model would model the probability of A being chosen as the best in the set (ABCDE), multiplied with the probability of B being chosen as the best in the subset (BCDE), then the probability of C being chosen as the best in the subset (CDE) and so on (Equation (1)).

Although the ROL ranks the implied order correctly, it ignores the way in which the data was obtained from the respondents: through alternating best-worst choices. That is why an alternative approach is proposed by Lancsar and Louviere (2008), the Sequential Best Worst Multinomial Logit (SBWMNL) model. In the same hypothetical case of five alternatives, alternative A would again be modelled as being chosen as the best among (ABCDE), but then multiplied with the probability of E being selected as the worst alternative among (BCDE), then B being the best of (BCD) and so on (Equation (2)).

Equation (1). *Rank Order Logit (ROL) model formulation* (Lancsar et al., 2013)

$$P\left(V_A > V_B > V_C > V_D > V_E\right) = \frac{e^{V_A}}{\sum_{j=A,B,C,D,E} e^{V_j}} \cdot \frac{e^{V_B}}{\sum_{j=B,C,D,E} e^{V_j}} \cdot \frac{e^{V_C}}{\sum_{j=C,D,\ E} e^{V_j}} \cdot \frac{e^{V_D}}{\sum_{j=D,E} e^{V_j}} \tag{1}$$

Equation (2). *Sequential Best Worst MNL model formulation* (Lancsar et al., 2013)

$$P\left(V_A > V_B > V_C > V_D > V_E\right) = \frac{e^{V_A}}{\sum_{j=A,B,C,D,E} e^{V_j}} \cdot \frac{e^{-V_E}}{\sum_{j=B,C,D,E} e^{-V_j}} \cdot \frac{e^{V_B}}{\sum_{j=B,C,D} e^{V_j}} \cdot \frac{e^{-V_D}}{\sum_{j=C,D} e^{-V_j}} \tag{2}$$

The difference between the methods is in the order in which the models are multiplied and which alternatives are removed from the 'remaining choice set' as well as that the SBWMNL model requires an assumption that the deterministic part of utility of choosing an alternative as worst is modelled as the negative of the deterministic utility of choosing that alternative as best.

2.2. *Image theory*

The goal of BWDCE is twofold. The first is to collect more information per choice set, compared to a standard 'choose the best' DCE. This requires that respondents need to process fewer choice tasks for the analyst to obtain a higher number of choice observations. The second is to make a difficult decision-making process easier for respondents, by having them alternatingly select the best/worst alternative. An important notion behind the use of BWDCE is that respondents find it relatively easier to identify the extremes within the choice set. A similar concept of decision-making can be observed in image theory (Beach and Mitchell, 1987), which states that consumers undertake a two-step process in decision-making when faced with multiple alternatives. First, they evaluate the alternatives based on their compatibility and exclude those that do not meet a subjective minimal standard. Once the choice set has been reduced to only include acceptable alternatives, these are evaluated on their profitability. According to Beach and Mitchell (1987), compatibility is a non-compensatory decision-making process, excluding alternatives that perform below a subjective threshold, while profitability is compensatory as consumers wish to maximise their 'profit' (utility) by trading-off the different characteristics (attributes) of alternatives. The definition of compatibility of Beach and Mitchell (1987) is similar to the decision-making process of Elimination by aspects (Tversky, 1972), where the decision-maker iteratively selects a certain attribute and removes all the alternatives that do not meet a certain aspect (a minimum performance on that attribute). Compatibility and profitability also draw strong parallels with RUM and P-RRM decision rules with respect to the level of compensatory behaviour.

The first step of the decision-making process in image theory was expanded upon by Ordóñez et al. (1999), stating that consumers have a certain subjective threshold and for each alternative, they add up its violations with respect to an expected performance. If the sum of the violations exceeds the threshold, the alternative is excluded from the second evaluation stage. As is shown in the following Section 2.3, this is reminiscent of the P-RRM model formulation, where only worse performance is considered and added up (a difference is that in an RRM model, the reference points are not expectations, but the performance of competing alternatives). Meloy and Russo (2004) added to the understanding of image theory by concluding that people prefer to choose by selection rather than rejection, as is the case in the second stage of image theory. They also found that the first stage does not seem to be the same in all cases. Exclusion of under-performing alternatives was found to be dominant in difficult situations, where a single correct answer is expected; this is in line with the notion that regret minimisation plays a larger role in decision making when choices are considered (more) difficult by the decision maker (Zeelenberg and Pieters, 2007). Conversely, in a setting where more subjective judgement needs to be passed, inclusion rather than exclusion seems to be the dominant thought process of narrowing down a choice set.

These findings motivate the hypothesis that respondents utilise different decision-making strategies when selecting best and worst alternatives. Behavioural theory suggests that exclusion of worst alternatives would be accomplished with non-compensatory decision rules, therefore a (P-)RRM decision rule can be expected to represent worst choices (stages 2 and 4). Best choices (stages 1 and 3) on the other hand can be assumed to be selected with the help of a RUM decision rule, as the consumer wishes to maximise their utility.

2.3. *Exploring (non-)compensatory behaviour with an RRM model*

The SBWMNL model as specified by Lancsar et al. (2013) is based on the linear in parameter random utility maximisation (RUM) paradigm. The utility of each alternative is based on the contribution of its individual components, as shown in Equation (3). This specification makes the RUM model fully compensatory, meaning that a below average performance on one attribute can be fully compensated by an above average performance on another attribute.

Equation (3). *RUM model formulation of systematic utility* (Ben-Akiva and Lerman, 1994)

$$V_i = \sum_{k \in K} \beta_k \cdot x_{ik} \tag{3}$$

As described in the previous subchapter, image theory postulates that some choices are made with a fully paradigm, while others with a non-compensatory behaviour. To analyse if the level of compensatory decision-making in BWDCE differs for different stages, a model that can capture such potential asymmetry is needed. This can be done by utilising the μRRM model, a generalised version of the RRM model with a regret parameter μ that varies between 0 and +∞ to capture everything between non-compensatory and fully compensatory behaviour (Van Cranenburgh, Guevara, et al., 2015).

Where RUM models consider only the performance of the alternative itself, RRM models look at the differences between the performances of alternatives on attributes. RRM models capture semi-compensatory decision-making behaviour, whereby a marginally larger penalty (compared to linear in parameter RUM models) is given for performing worse than another alternative, while a marginally smaller bonus is given for performing better. The result is a semi-compensatory model which postulates that the extent to which a poor performance on one attribute can be compensated by a strong performance on another, depends in subtle ways on the composition of the choice set (Chorus, 2010).

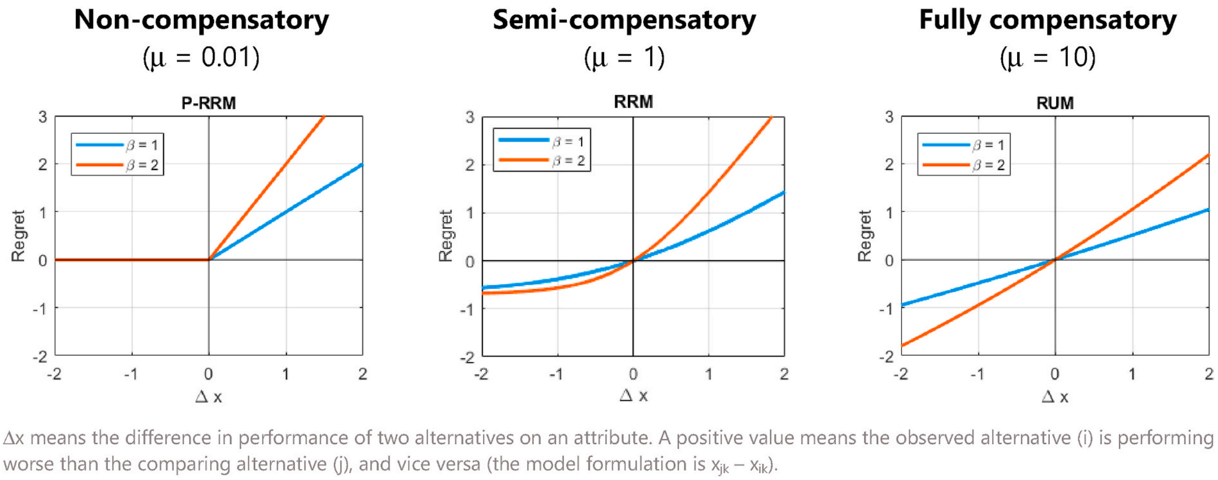
The μRRM model (Van Cranenburgh, Guevara, et al., 2015) is a generalisation of the RRM model proposed in by Chorus (2010); it includes an additional parameter μ, which allows the decision rule of the respondent's decision-making approach to be on a spectrum anywhere between non-compensatory behaviour and fully compensatory behaviour. The scale parameter (regret parameter) varies between values of 0 and +∞ which correspond to non- and fully-compensatory behaviour respectively. When behaviour is fully-compensatory, performing better yields the same rejoice (utility) as performing worse to an equal degree yields regret (disutility). In this case, the model produces the same outcome as a RUM model. If behaviour is non-compensatory however, it means that performing worse will still result in regret, but performing better will give no rejoice; this so-called P-RRM model (pure RRM) generates the most extreme choice set composition effects, such as a compromise effect which rewards alternatives that offer a compromise in the sense of having an intermediate performance on most attributes. Fig. 1 shows the different levels of regret associated with an alternative's performance given each of the three distinct cases of decision-making behaviour: non-, semi- and fully compensatory behaviour. The corresponding μ values used to illustrate the behaviour are shown in the figure as well. A value of 0.01 is used for non-compensatory because the parameter does not actually reach zero, so this approximation is used. For the fully compensatory decision rule, any value above ten already yields such behaviour and using a higher value yields not difference in the decision-making behaviour (Van Cranenburgh, Guevara, et al., 2015). The μ scale parameter in the μRRM model should not be confused with the μ scale parameter related to the variance of the error term in RUM models, defined as $\mathrm{var}(\varepsilon) = (\pi^2/6)/\mu^2$ and fixed to allow the estimation of taste parameters (Train, 2009). To avoid confusion between the two, we refer to the parameter in the μRRM model as the regret parameter from here on.

The model formulation of the RRM model differs from the RUM model in the way the alternatives' performance is evaluated. As shown in Equation (3), RUM uses the level of an attribute (x_i), the μRRM model considers the difference between attribute levels of different alternatives (x_{jk} - x_{ik}). The μRRM model (Equation (4)) extends the RRM model with the addition of the μ regret parameter. By doing so, this regret parameter can capture any level of compensatory behaviour, including fully compensatory behaviour found in the standard RUM model, as well as completely non-compensatory behaviour.

Equation (4). *μRRM model formulation* (Van Cranenburgh, Guevara, et al., 2015)

$$R_i = \sum_{j \neq i} \sum_{k=1 \ldots K} \mu \cdot \ln\left(1 + \exp\left(\frac{\beta_k}{\mu} \cdot (x_{jk} - x_{ik})\right)\right) \tag{4}$$

Due to the nature of their model specification, μRRM models rely on the comparison multiple alternatives and attribute levels to



Δx means the difference in performance of two alternatives on an attribute. A positive value means the observed alternative (i) is performing worse than the comparing alternative (j), and vice versa (the model formulation is $x_{jk} - x_{ik}$).

Fig. 1. Levels of compensatory behaviour.

obtain the regret of an alternative. To enable the estimation of a μRRM model that produces different results than a RUM model, at least three alternatives need to be present in the choice set and attributes need to have at least three levels. In cases when only two alternatives are present, the model is indistinguishable from a RUM model and the value of the regret parameter in the μRRM model has no impact on the outcome of the model estimation (Van Cranenburgh, Guevara, et al., 2015). As μRRM models rely on comparing attribute levels between alternatives, the alternatives should have as many generic attributes as possible (ideally all) – attributes that are present in all alternatives.

To accommodate the use of the μRRM model in the SBWMNL to determine the regret/utility of individual alternatives, a few changes need to be made to the model formulation. Firstly, the signs need to be adapted to accommodate the change from a RUM to an RRM model. When modelling first-best choice experiments with an RRM MNL model, the negative value of regret (-*R*) is used in the logit function to determine the choice probability of the selected alternative. The SBWMNL already has both positive and negative signs in the exponent to accommodate the best-worst decision making behaviour. Using a negative sign in the exponent allows the calculation of the alternative's probability of being chosen as the worst alternative, as the probability of being selected as worst is the inverse of being selected as the best, as discussed by Lancsar et al. (2013). Therefore, to adapt the SBWMNL for the μRRM model, the signs need to be reversed from + to – and vice-versa. Hence a negative sign is present when modelling alternatives selected as "best", and a positive sign when modelling alternatives selected as "worst" (Equation (5)).

Equation (5). *Formulation of the SBWMNL μRRM model*

$$P\left(R_A < R_B < R_C < R_D < R_E\right) = \frac{e^{-R_A}}{\sum_{j=A,B,C,D,E} e^{-R_j}} \cdot \frac{e^{R_E}}{\sum_{j=B,C,D,E} e^{R_j}} \cdot \frac{e^{-R_B}}{\sum_{j=B,C,D} e^{-R_j}} \cdot \frac{e^{R_D}}{\sum_{j=C,D} e^{R_j}} \tag{5}$$

Secondly, unlike RUM models, μRRM models are choice-set-size dependent, meaning that the number of alternatives in a choice set influences the scale of regret for each alternative and thus comparisons between differently sized choice sets are not possible (Guevara et al., 2016). This arises from the μRRM model formulation, as the scale of regret comes from a summation of comparisons to other alternatives and a larger choice set will result in more comparisons being made. This could prove problematic in an SBWMNL model, where the variation of choice set size is essential to the methodology. Van Cranenburgh et al. (2015) offer two ways of dealing with choice set size variation. The first, applicable in particular when choice set sizes are known in advance, is to multiply the regret of each alternative with a correction factor (Λ_j), specific to the choice set size (J). In a BWDCE with five alternatives, three correction factors are needed, as the choice set size varies between five and two alternatives and a correction factor needs to be added to all but one choice set size. In the second approach, the regret of each alternative is still multiplied with a correction factor, but that factor is made of the choice set size (J_n) and a single constant (Λ). Using the same example, the correction factors would thus be 5/Λ, 4/Λ, 3/Λ and 2/Λ. The second approach is seen as more flexible, as it gives the possibility to apply it to virtually any size choice set, although the reliability of extrapolation to sizes not estimated in the model is not guaranteed. A downside of this approach is that it cannot capture the scale variability in the different stages as well as the first approach, using choice set size specific correction factors.

2.4. Model estimation

To analyse whether respondents do indeed use different levels of compensatory behaviour for different stages, a two-step approach

Table 1
Overview of estimated models.

Group	Model name	Model formulation	Modelled stage (s)	Number of μ param.	Number of Λ param.	Comments
Benchmark models	B1	RUM	1	/	/	RUM MNL model
	B2	RUM	all	/	/	SBWMNL model
	B3	RUM	all	/	3	Scaled best-worst MNL model
Implied decision rule models	UU	μRRM	all	2 (fixed)	3	μ_1 & 3 = *10,* μ_2 & 4 = *10*
	RR	μRRM	all	2 (fixed)	3	μ_1 & 3 = *0.01,* μ_2 & 4 = *0.01*
	UR	μRRM	all	2 (fixed)	3	μ_1 & 3 = *10,* μ_2 & 4 = *0.01*
	RU	μRRM	all	2 (fixed)	3	μ_1 & 3 = *0.01,* μ_2 & 4 = *10*
Estimated decision rule models	E1	μRRM	all	1	3	
	E2	μRRM	all	2	3	
	E3	μRRM	all	4	3	
	E4	μRRM	all	1	1	
	E5	μRRM	all	2	1	
	E6	μRRM	all	4	1	
Individual stage models	S1	μRRM	1	1	/	
	S2	μRRM	2	1	/	
	S3	μRRM	3	1	/	
	S4	μRRM	4	1	/	

All models have 5 taste (marginal utility) parameters, that always cover all the modelled stages.

is taken in the modelling process. Firstly, four models are estimated in which the regret parameters determining the decision rules are fixed to represent either fully or non-compensatory behaviour. The four models cover all four combinations of best and worst choices with fully and non-compensatory behaviour (Table 1). As fully compensatory behaviour is synonymous with linear in parameter RUM models, whereas non-compensatory behaviour is with RRM models, the models are named to reflect that: in the UR model for example, fully compensatory behaviour is assumed for best choices (stages 1 and 3) and non-compensatory for worst choices (stages 2 and 4).

Secondly, regret parameters in the SBWMNL μRRM are estimated, to determine what is the extent of compensatory behaviour in the decision-making process. Six models are specified (Table 1), where the number of regret parameters is varied and both approaches to modelling choice set size variation are applied. As the SBWMNL RUM model assumes a fully compensatory behaviour, models E1 and E4 are setup as a direct comparison, with a single regret parameter for all four choices/stages in the decision-making process to test that if a single decision rule is assumed for all choices in the BWDCE, is a fully compensatory decision rule a valid choice.

The remaining four models (E2, E3, E5 and E6) cover the four combinations of either two or four regret parameters and the two choice-set-size-variation mitigation approaches. In models where two regret parameters are used (E2 and E3), one is used for Stages 1 and 3 of the decision-making process, where the respondents have to pick the ‘Best’ alternative, whereas the other regret parameter is for choices in Stages 2 and 4, where the ‘Worst’ alternative has to be selected. In models E5 and E6, four different regret parameters are estimated, one for each stage of the decision-making process.

To compare the outcomes of the developed model with existing models, three benchmark models, are estimated (Table 1): first-choice-only RUM MNL model (B1), the SBWMNL RUM model developed by Lancsar et al. (2013) (B2) and an SBWMNL RUM model with three scale parameters (B3) accounting for the four stages of the decision-making process. The latter model is included to make a more fair comparison with models E1-3, which also utilise three scale parameters accounting for the different scale of the individual stage in the BWDCE.

In all the aforementioned models (E1-6 and B1-3), a single set of taste (marginal utility) parameters is estimated, meaning that the same parameters are used for all stages of the choice experiment.

Additionally, stage-specific models (Table 1) are also estimated for all four stages (S1–S4). This is done to test the stability of successive choices in the same choice set, decouple the individual stages and analyse the model fit and parameter estimates of each stage independently.

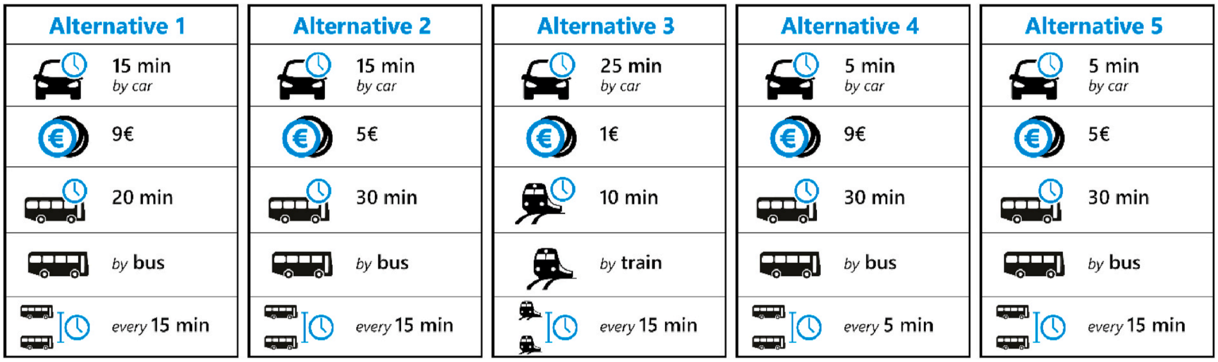
2.5. *Statistical tests*

Statistical tests are used to evaluate and compare the performance of the above-listed models. The Likelihood Ratio Test (LRT) can be used to compare nested models (Hauser, 2008) and can in our research be used to compare the E models. For comparing non-nested models (the implied decision rule models), the Ben-Akiva & Swait test is used (Ben-Akiva and Swait, 1986). Both statistical tests compare the model fit with respect to the number of parameters used in the model estimation and are used to evaluate model parsimony (to avoid overfitting). The outcome of both tests is the probability that the worse fitting model (model with the lower fit) is, despite its lower model fit, the true model for the population. To further assess the parsimony of the models, their BIC value (Stone, 1979) is also reported.

2.6. *Data collection*

The data used to estimate the developed models were collected specifically for the purpose of this study. The choice experiment is designed in a way to allow the estimation of an SBWMNL μRRM model. For the BWDCE, more than two alternatives (and preferably even more) are needed to arrive at an implied ranking based on best-worst data. In line with what Lancsar et al. (2013) have done, the choice sets are designed with five alternatives, which means that in each choice set, respondents have to evaluate four stages (choices) in order to obtain an implied ranking of alternatives: the four choices are BEST-WORST-BEST-WORST.

For the RRM part of the model, there are two additional requirements, as explained in Section 2.3: attributes need to have at least three levels and they need to be generic. To accommodate these requirements, the topic of the survey was chosen to be ‘Park and Ride



Alternative 1	Alternative 2	Alternative 3	Alternative 4	Alternative 5
15 min *by car*	15 min *by car*	25 min *by car*	5 min *by car*	5 min *by car*
9€	5€	1€	9€	5€
20 min	30 min	10 min	30 min	30 min
by bus	*by* bus	*by* train	*by* bus	*by* bus
every 15 min	*every* 15 min	*every* 15 min	*every* 5 min	*every* 15 min

Fig. 2. Example of a (translated) choice set presented to the respondents.

(P + R) facility choice' and had 12 choice sets, each made of five unlabelled alternatives (Fig. 2) with five generic attributes (with three levels), namely:

- Access time to the P + R facility by car (5, 15, 25 min)
- Combined cost including parking and a public transport ticket (1, 5, 9 €)
- Travel time by public transport (10, 20, 30 min)
- Public transport mode (bus, train)
- Public transport frequency (5, 15, 30 min)

The experimental design is constructed with a Bayesian efficient design using priors obtained from Bos et al. (2004), who conducted a P + R research in the Netherlands. A Bayesian efficient design is chosen over a regular D-efficient design due to uncertainty in the transfer of context between Netherlands and Slovenia, and as Walker et al. (2018) point out, the efficiency of a D-efficient design can drop if the actual value of time differs significantly from the value of time used in the priors.

In designing the survey, we used a novel approach proposed by van Cranenburgh, Rose and Chorus (2018), as they show that traditional efficient designs evoke behaviour that is in line with RUM decision rules. The design they propose is therefore efficient not only for estimating RUM model-based parameters, but also for estimating RRM model-based parameters. Since this study had been carried out, this survey design approach had also been incorporated into the Ngene software (van Cranenburgh and Collins, 2019).

The survey was conducted online, among residents of and daily commuters to the city of Ljubljana, Slovenia in June and July of 2018. In total, 108 complete responses were recorded and 138 partial responses. The survey was deemed completed if all the stated choice questions were answered. The survey also included questions on the respondents' socio-demographics and travel behaviour, but as these were not crucial for the model estimation, a response was considered complete even if those questions were left unanswered. With each of the 108 respondents having evaluated 12 choice sets, in each of which four choice were made, this gave us a total of 5184 observations.

3. Results

The section starts with the *Implied decision rule* models in Section 3.1, analysing only the model fit with respect to implying either a fully or a non-compensatory decision rule for selecting best/worst alternatives. Secondly, the *Estimated decision rule* models are compared to the benchmark models and the estimated regret parameters are analysed in Section 3.2. Thirdly, in Section 3.3, the stability of the individual stages of the decision-making process are evaluated. Complete model results and a comparison of parameter ratios are presented in the Appendix.

3.1. Implied decision rule models

The model fits of the four models with implied decision rule combinations (Table 2) show that models UU and UR – the models with a fully compensatory decision rule for 'best' choices – significantly outperform models RR and RU, which assume non-compensatory behaviour in selecting best alternatives, with a difference in rho-squared of more than 0.02. The Ben-Akiva & Swait test (Ben-Akiva and Swait, 1986) shows that this difference is highly significant, meaning that the probability of models RR or RU being the true/correct models for the population is close to zero.

The difference between the performance of models UU and UR (estimating a different decision rule for choosing worst alternatives) is however less pronounced and equals 4.9 LL-points. According to the Ben-Akiva & Swait test however, given the number of observations, this difference in Final Log-Likelihood is significantly different. The less fitting UR model only has a 0.00087 probability of being the true model for the population.

These results are partially in line with image theory, as they state that a fully compensatory decision rule is more likely to be used when choosing the best alternative. The results also show that in this particular case, the worst alternatives were not chosen using a non-compensatory decision rule. The following subsection explores what value the regret parameter (μ) takes when it is estimated in the model along with the taste parameters.

The full model results, including the estimated taste parameters can be found in Appendix A in Table 5.

3.2. Estimated decision rule models

Comparing the estimated E models to the benchmark model B2 (the SBWMNL model), all six of them outperform the B2 model in terms of model fit (Table 3) and according to the Ben-Akiva & Swait test, the differences in model fit are all highly significant.

Considering the B3 benchmark model, it outperformed four of the six estimated models, namely models E1, E4, E5 and E6. This is largely due to the three additional scale parameters that account for the scale difference in the successive stages of the choice task. Considering the Ben-Akiva & Swait test, the B3 model significantly outperforms the four E models. For models E4-E6, this is less surprising as they have only a single parameter accounting for choice set size and can therefore not accommodate the scale variability in the same way.

The difference between models B3 and E1 on the other hand is very small, with the Ben-Akiva & Swait test predicting that E1 (although with a slightly lower model fit), has a 0.034 probability of being the true model in the population, compared to B3. The difference in Log-Likelihood is only 0.16. The two models use different specifications (RUM or μRRM), but as the regret parameter in

Table 2
Model fits of models with fixed regret parameters.

	UU	RR	UR	RU
Observations	5184	5184	5184	5184
Parameters	8	8	8	8
Null LL	−6204.59	−6204.59	−6204.59	−6204.59
Final LL	−5346.00	−5488.43	−5350.90	−5486.91
Rho-squared	0.1384	0.1154	0.1376	0.1157
BIC	10,760.42	11,045.29	10,770.22	11,042.25

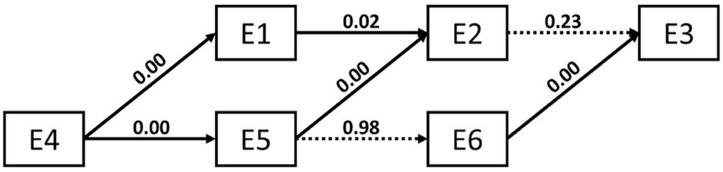
Table 3
Model fit results and parameter estimates, with parameter estimate t-values in brackets.

	B2	B3	E1	E2	E3	E4	E5	E6
Observations	5184	5184	5184	5184	5184	5184	5184	5184
Parameters	5	8	9	10	11	7	8	9
Null LL	−6204.59	−6204.59	−6204.59	−6204.59	−6204.59	−6204.59	−6204.59	−6204.59
Final LL	−5397.20	−5345.84	−5346.00	−5340.49	−5337.50	−5382.88	−5372.99	−5372.95
Rho-squared	0.1301	0.1384	0.1384	0.1393	0.1398	0.1324	0.1340	0.1340
BIC	10,837.17	10,760.10	10,768.97	10,766.51	10,769.08	10,825.63	10,814.40	10,822.88
Model formulation	RUM	RUM	μRRM	μRRM	μRRM	μRRM	μRRM	μRRM
Regret parameters								
μ_all			10 (fixed)			10 (fixed)		
μ_1 & 3				10 (fixed)			10 (fixed)	
μ_2 & 4				0.43 (2.54)			0.36 (2.88)	
μ_1					10 (fixed)			10 (fixed)
μ_2					0.43 (2.56)			0.36 (2.88)
μ_3					0.50 (1.97)			12.31 (0.6)
μ_4					1 (fixed)			1 (fixed)
Scale parameters								
Λ_all						5.49 (8.11)	5.23 (7.87)	5.25 (7.88)
Λ_5		2.13 (7.97)	0.86 (fixed)	0.87 (7.87)	0.88 (7.79)			
Λ_4		1.20 (7.17)	0.61 (12.54)	0.70 (6.93)	0.71 (6.87)			
Λ_3		1.48 (7.53)	1.00 (14.72)	1.00 (7.44)	1.11 (7.09)			
Λ_2		1 (fixed)	1 (fixed)	1 (fixed)	1 (fixed)			
Taste parameters								
Car in-vehicle time	−0.0581	−0.0372	−0.0371	−0.0367	−0.0366	−0.0374	−0.0368	−0.0369
	(−25.51)	(−8.30)	(−23.16)	(−8.15)	(−8.07)	(−8.28)	(−8.06)	(−8.08)
PT in-vehicle time	−0.0453	−0.0282	−0.0282	−0.0282	−0.0282	−0.0282	−0.028	−0.0281
	(−20.02)	(−7.99)	(−19.5)	(−7.86)	(−7.8)	(−7.94)	(−7.77)	(−7.79)
Cost	−0.2274	−0.1414	−0.1412	−0.1399	−0.1392	−0.1386	−0.1367	−0.1371
	(−33.43)	(−8.27)	(−27.6)	(−8.1)	(−8.02)	(−8.19)	(−7.98)	(−8.01)
PT headway	−0.0383	−0.0259	−0.026	−0.0264	−0.0265	−0.0259	−0.0267	−0.0268
	(−19.08)	(−8.02)	(−17.57)	(−7.91)	(−7.86)	(−7.95)	(−7.86)	(−7.88)
PT mode	0.1864 (5.18)	0.099 (3.80)	0.1004 (4.57)	0.1023 (3.97)	0.1037 (4.08)	0.0933 (3.61)	0.0954 (3.76)	0.0957 (3.76)

E1 indicates a fully compensatory decision rule, it means that both models assume the same single decision rule throughout the stages of the choice task.

Comparing B3 to E2 and E3, the latter two prove to be significantly better, with the Ben-Akiva & Swait test resulting in a 0.015 and 0.003 probability respectively, of being the true model for the population.

This shows that there seems to be some variation present in the decision-making process, that the regret parameters in the E models are able to capture and better represent.



The arrow between two models indicates nesting, allowing for a Likelihood Ratio Test. The model to the left of the arrow has a lower fit, and the p-value on the arrow indicates the probability of the worse fitting model being the true model in the population, compared to the better fitting model located to the right of the arrow.

Fig. 3. Likelihood ratio tests among the E models.

Considering benchmark model B1 (model outcomes presented in Table 5 in the Appendix), although it achieved a model fit (rho-squared) twice as high as all the other models, this comparison is not appropriate. Model B1 is a traditional RUM MNL model, that uses only the choices made in the first stage of the choice process, which in this case consists only of a quarter of the observations. The higher performance of the first stage is discussed further in Section 3.3.

Evaluating the significance in the difference of their model fit with the LRT, it can be seen in Fig. 3 that five out of seven pair-wise comparisons are significant, meaning that those models are significantly different. This cannot be said for the models in the two pairs (E2-E3 and E5-E6) where the difference in model fit cannot be said to be significant.

Model E3 has the largest number of parameters and thus performed the best in terms of rho-squared, followed closely by model E2. The LRT reveals that E3 is not significantly better than E2 and the lower BIC value of model E2 indicates that the additional parameters in model E3 do not yield a sufficient increase in model fit to justify their use. The same can be said about models E4 and E5. What both of these pairs have in common is that models E3 and E6 expanded on the number of regret parameters from E2 and E5 respectively, from two to four, indicating that in both cases, having a separate regret parameter for each stage of the choice task does not yield a significantly better result than only having two regret parameters, one for the best-choice stages and one for the worst-choice stages.

Increasing the number of regret parameters from one to two, as in the model pairs E4-E5 and E1-E2 however, does yield a sufficient increase in model fit. This means that adding a second regret parameter in models E2 and E5 to separately model stages 2 and 4 (where respondents had to choose the worst alternative) does seem to be advantageous.

Turning to the values of the regret parameters in the different models, as seen in Table 3, the findings from the previous subsection are confirmed, as the regret parameters for the first and third stages (modelling best choices) mostly take a value that is equivalent to fully compensatory behaviour. As anticipated the regret parameters used to model the second and fourth stages (choice of worst alternative) indicate a decision-making behaviour that is somewhere between fully and a non-compensatory.

Looking first at the stages 1 and 3 ("best" choices), the decision rule in the first stage in particular seems to be made using a fully compensatory decision rule, with all the regret parameters taking a value of 10.[1] In models E3 and E6, where a separate parameter was used to model the third stage (μ_3), it once converged to a value of 0.5 (semi- to non-compensatory) and once to 12.31 (fully compensatory), however both estimates were insignificantly different from a value of one. The difference in parameter estimates can be explained by the different modelling approach that was used to account for choice set size variation.

A similar confirmation from the previous section can also be seen for choices made in stages 2 and 4 ("worst" choices), in that they seem to be made with a fairly strong semi- to non-compensatory behaviour. The estimated values of 0.43 and 0.36 for regret parameter μ_2 & 4 in Table 3 mean that some trading-off is being made, but to a very limited extent, and that performing better yields only a limited amount of rejoice, compared to the level of regret for the same difference in performance, which is (again) close to how image theory describes decision-making behaviour. Parameter μ_4 in models E3 and E6 was fixed to an arbitrary value of 1[2].

Finally, in models E1 and E4, where a single regret parameter value is used for all four stages of the choice process, the decision rule does seem to be fully compensatory. The use of the SBWMNL RUM model (Lancsar et al., 2013) is therefore the most appropriate model to use on our data, if a common decision rule is assumed for all stages of the choice process. This result also points to an interesting finding of how the regret parameter values change when reducing the number of stages they cover. It seems that the decision rules of later-stage choices are less important, as they always yield to the earlier-stage choices. This is due to the lower stability of later-stage choices and is discussed further in the following subsection.

With respect to the taste parameters in different models, evaluating the parameter ratios shows there is very limited difference between models. The implications of this are discussed in more detail in Section 4 and a detailed overview of the significance of parameter ratio differences is presented in Appendix B.

3.3. Stability of data collected with BW experiments

As shown in the previous sections, the first decision stage (first-best) seems to dominate all other stages in cases when they have a common regret parameter. In E1, the regret parameter for all four choices is estimated to be fully compensatory. Then in E2, when the second and fourth stage are given a separate parameter, it shows almost the completely opposite decision rule. Then in E3, the same happens with the third stage choice (second-best), as it also moves towards non-compensatory behaviour. This seems to point towards a higher importance and stability of the first-stage choice and its impact on the final LL, compared to the later-stage parameters which seem to be more unstable. This is found to be the case, when the values of regret parameters in model E3 are varied one at a time between 0,01 (non-compensatory) and 10 (fully compensatory). Fig. 4 shows how the final LL changes as a result of changing the regret parameter, and the difference in final LL between a non-compensatory and a fully compensatory regret parameter decreases from 50 LL-points in the first stage, to only 5 LL-points in the second and a mere 3 LL-points in stage three. This explains why, when stages are given a common regret parameter, they are all influenced by the behaviour observed in the first stage.

[1] Regret parameters had to be fixed to a value of 10 to allow the model estimation process to converge. Although the regret parameter varies between zero and infinity, any value above 10 can be considered to represent fully compensatory decision-making and this can cause problems in the model estimation process, where values above 10 all give essentially the same model fit and the model estimation cannot converge (Van Cranenburgh, Guevara, et al., 2015). The regret parameters in each model are therefore fixed one at a time, until the model estimation converged. The non-converged model estimates for the regret parameter μ_1 & 3 and μ_1 ranged between 12 and 60.

[2] In these choice sets, only two alternatives were present, meaning that RUM and RRM models are indistinguishable from each other and the model would not converge, as any value of μ_4 would yield the same outcome (Van Cranenburgh, Guevara, et al., 2015).

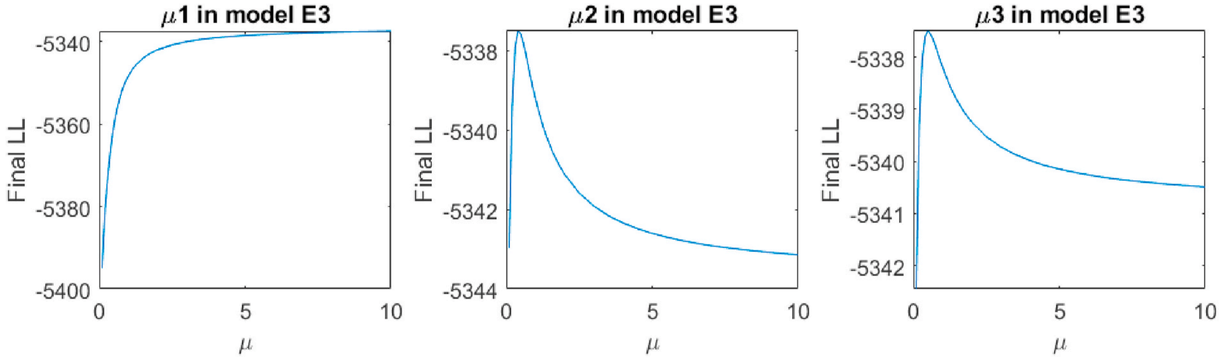


Fig. 4. Change in final LL resulting from the variation of the regret parameter.

The issue of decreasing stability of successive stages is also observed when estimating the models for each stage separately, with independent taste and regret parameters (note that in all the E models, even when the regret parameters are stage-specific, the taste parameters are always equal for all stages). By doing so, a higher model fit of the first stage is revealed. Table 4 shows how the rho-squared value decreases following the first stage. These findings are in line with what other researchers reported on the stability of choice made in the later stages (Ben-Akiva et al., 1992; Dyachenko et al., 2014; Giergiczny et al., 2017). The notion that a choice which is made in the first stage, containing a larger choice set, offers more information in the process of parameter estimation (in an econometric as well as information theoretic sense) than a choice which is made in a later stage, is a likely (partial) explanation for this phenomenon.

4. Discussion & conclusion

Best-worst data collection techniques have become popular in recent years, due to their ability to capture a larger number of observations in smaller samples, enabling discrete choice analyses to be carried out for smaller populations/samples and allowing researchers with a smaller budget to obtain sufficient observations for discrete choice model estimation. Use of this data collection technique raises the question of decision rule consistency throughout the experiment. In other words, do the respondents utilise the same (fully compensatory) decision rule throughout the experiment, or do they switch between fully and non-compensatory behaviour, as Beach and Mitchell (1987) defined decision-making behaviour in image theory.

Our study shows that models accounting for different decision rules being used throughout the BWDCE tend to perform better than models that assume a single decision rule throughout the choice process. The decision rules emerging from these models are to a large extent in line with what image theory postulates. While we cannot say for certain that the variability of the decision rule is what is driving this difference in model performance, we do observe that the best performing models result in a fully compensatory decision rule for stages where respondents have to select the best alternative and a semi- to non-compensatory decision rule for stages where they have to select the worst performing alternative.

A notable exception in results with respect to image theory are the choices made in the third stage (the second time choosing the best alternative), where a fully compensatory decision rule is expected, but model estimates show a semi- to non-compensatory decision rule. This is the same kind of decision rule as estimated for the second and fourth stage (worst choices). This only happened when the third stage is modelled separately from the first stage, with its own regret parameter, as in the case of a joint regret parameter for both the first and third stages (first and second best choices) the former seems to overpowered the latter and a fully compensatory decision rule estimate is obtained. This raises the issue of lower stability in successive choices (Ben-Akiva et al., 1992; Dyachenko et al., 2014; Giergiczny et al., 2017), as the additional choices people make after selecting their preferred alternative seem to be less important to them. We investigated the stability of the separate stages in our data as well and came to the same conclusion as Ben-Akiva et al. (1992) and Giergiczny et al. (2017), with the first stage achieving a model fit that is between two and four times higher than any of the successive choices.

One reason for the lower reliability of successive stages could be that after selecting the preferred alternative in the first stage, respondents may remember the performance of said alternative and compare all the others to it. The models do not capture this, as

Table 4
Model fit results of the different stages of the choice task.

	S1	S2	S3	S4
Observations	1296	1296	1296	1296
Parameters	6	6	6	6
Null LL	−2085.83	−1796.64	−1423.80	−898.32
Final LL	−1548.82	−1669.14	−1251.02	−841.53
Rho-squared	0.2575	0.0710	0.1213	0.0632

each stage is modelled based solely on the alternatives presented in the choice set. After an alternative was selected (either as best or worst), it was removed from the choice set in the next decision stage.

Another reason behind the lower reliability of successive stages could be the rather unnatural way of decision-making, as normally we do not keep evaluating our options after we have already made our decision, especially not by choosing a single alternative we deem as the least appropriate. This unnatural feeling was also pointed out by some of the respondents, stating that they found it confusing. The stage-specific model outcomes are in line with these concerns and interestingly also uncover a lower stability of the second and fourth stage (worst choices), compared to the third stage (second best choice), showcasing that worst-choice stages may be more unnatural than successively choosing the best alternative. The third stage, where respondents choose the (second) best alternative is perhaps seen as a back-up choice if the first preference is unavailable and thus given more thought by the decision-makers. These findings are in line with Dyachenko et al. (2014), who warn that particularly in fields where typical market conditions apply - to which the topic of daily commuting in our survey adheres – the use of BW choice tasks should be avoided, as it is an unnatural decision-making process. Ghijben, Lancsar and Zavarsek (2014) note that BW tasks require the respondent to 'swap to a new mental task' when they switch between choosing the best and worst alternatives (Lancsar et al., 2017, p. 705). For cases where limited samples are available and multiple observations per choice set would still be advantageous, Ghijben et al. (2014) propose the use of a best-best (BB) DCE instead. Although not immune to the lower stability of choices in successive stages, our results also confirm that the stability of successive best choices is higher than that of worst choices. The majority of researchers are clear, that if a large enough sample can be obtained (a large enough population exists), any kind of successive choices should be avoided. That is why BW and BB choice tasks are more prominent in the field of healthcare, where population and sample sizes can often be limited, whereas the same can typically not be said for the field of transportation.

It is also important to briefly discuss the parameter estimates originating from the different models. The comparison of parameter ratios (Appendix B) shows that most models, despite having vastly different formulations and modelling different decision rules, produce very similar parameter ratios, with only a handful showing significant differences. While this may suggest that only analysing the first stage of the BWDCE is sufficient, we emphasise that BWDCEs are meant for small sample sizes. Our results show that the dataset of 1296 observations (from 108 respondents) is already sufficient to analyse only using first stage choices. It does also showcase, that modelling all the choices made in the BWDCE does not significantly change the model estimates outcome and could thus prove advantageous in circumstances where a large sample of respondents for first-stage-only choices cannot be obtained.

In addition to the BW decision process, using RRM models comes with its own set of downsides, most notably its inability to determine the net welfare effect (the value of a choice set). Dekker (2014) and more recently Dekker and Chorus (2018) proposed new metrics for RRM models, giving them more applicability in (transportation) policy evaluation, but the authors still acknowledge the limitations of the RRM model in welfare analyses when compared to linear in parameter RUM models. In addition to this, using the SBWMNL μRRM model for forecasting is problematic due to different choice set sizes, which have to be accounted for in RRM models. Especially if the forecasting will be made for choice set sizes not used in the estimation, a single correction factor adjusted for choice set size, needs to be used. This approach is more flexible and can be applied to choice set sizes not included in the model estimation, although the extrapolation to choice set sizes beyond those in the estimation process is questionable. Due to its more flexible nature, the approach is also less reliable, results in a lower model fit and can potentially influence the value of the regret parameter to represent a completely different level of compensatory behaviour.

Despite the issues facing the use of successive-choice models and RRM models, our results still provide valuable insight into the behaviour of respondents in BWDCE, giving researchers who use the method a good understanding of how their respondents behave, how some of their actions can be explained and how the decision-making process looks in such an experimental setting. It also gives them a tool to test the decision-making behaviour of their respondents, before making any assumptions on their behaviour.

Chorus et al. (2014) found that neither a fully nor non-compensatory decision rule is used in a majority of cases by respondents in traditional SP experiments (first-best only). This means that also in BW choice tasks, no fixed decision-rule combination can be assumed to be utilised by respondents, and the use of either a fully or non-compensatory decision rule can vary significantly between different samples due to a variety of reasons, such as the context/topic of the survey, decision-maker groups, different cultures etc. and thus no assumptions can be made on any given sample prior to the analysis of the results. To accurately carry out an analysis of BW data, the sample should be tested for the decision-rule combination used by the respondents and any simplifications to the model can follow thereafter, based on the uncovered decision-making behaviour.

BWDCE also offer a variety of avenues to be researched further. Further application of our SBWMNL μRRM model could help paint a clearer picture of what is the more/most common decision-rule combination used by respondents or if, similar to the findings of Chorus et al. (2014), there is no clear preference and a number of different combinations are equally likely to occur. This research also used a novel approach to create the experimental design that made no assumptions on the underlying decision rule (Van Cranenburgh et al., 2018), but this approach (or any other approach for generating efficient experimental designs) does not consider that respondents make multiple choices in the same choice set. Such an experimental design methodology would be beneficial given the rise of BWDCE use. Although our model found differences in decision rules between the stages, the model is limited to the μRRM formulation which only considers the range of fully to non-compensatory decision rules and testing the decision rule heterogeneity with other models that are able to capture other decision rules could offer further insights into respondent behaviour.

Declaration of interest

There is no conflict of interest for any of the five contributing authors.

CRediT authorship contribution statement

Nejc Geržinič: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Sander van Cranenburgh:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Oded Cats:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Emily Lancsar:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Caspar Chorus:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Appendix A. Model estimation results of models evaluating a single stage in the choice task

Table 5
Model fit and parameter estimate results of the different stages of the choice task (t-val in brackets)

	B1	S1	S2	S3	S4
Observations	1296	1296	1296	1296	1296
Parameters	5	6	6	6	6
Null LL	−2085.83	−2085.83	−1796.64	−1423.80	−898.32
Final LL	−1549.08	−1548.82	−1669.14	−1251.02	−841.53
Rho-squared	0.2573	0.2575	0.0710	0.1213	0.0632
Modelled stage	1	1	2	3	4
Model formulation	RUM	µRRM	µRRM	µRRM	µRRM
M		53.5 (27120.53)	0.54 (1.70)	0.67 (2.00)	1 (fixed)
Car in-vehicle time	−0.0874 (−22.73)	−0.035 (−252.16)	−0.0195 (−7.00)	−0.0372 (−10.90)	−0.0285 (−5.02)
PT in-vehicle time	−0.0601 (−16.82)	−0.0241 (−188.77)	−0.0235 (−8.17)	−0.03 (−8.68)	−0.0265 (−4.62)
Cost	−0.2918 (−24.22)	−0.1168 (−542.76)	−0.0934 (−12.53)	−0.1617 (−14.4)	−0.1706 (−9.87)
PT headway	−0.0643 (−16.02)	−0.0257 (−245.87)	−0.0135 (−5.31)	−0.0277 (−9.06)	−0.0245 (−5.18)
PT mode	0.1875 (3.09)	0.0770 (216.09)	0.0017 (0.04)	0.1378 (2.75)	0.2942 (3.25)

Table 6
Model fit and parameter estimate results of the different stages of the choice task (t-val in brackets)

	UU	RR	UR	RU
Observations	5184	5184	5184	5184
Parameters	8	8	8	8
Null LL	−6204.59	−6204.59	−6204.59	−6204.59
Final LL	−5346.00	−5488.43	−5350.90	−5486.91
Rho-squared	0.1384	0.1154	0.1376	0.1157
BIC	10,760.42	11,045.29	10,770.22	11,042.25
Car in-vehicle time	−0.0372 (−23.16)	−0.0321 (−24.74)	−0.0363 (−23.43)	−0.033 (−22.78)
PT in-vehicle time	−0.0282 (−19.5)	−0.027 (−20.29)	−0.0281 (−20.17)	−0.0273 (−18.64)
Cost	−0.1414 (−27.61)	−0.1311 (−28.18)	−0.1387 (−27.75)	−0.1348 (−24.58)
PT headway	−0.0261 (−17.57)	−0.0276 (−21.29)	−0.0267 (−19.05)	−0.0277 (−18.34)
PT mode	0.1006 (4.57)	0.1144 (6.76)	0.1068 (5.16)	0.1104 (6.1)
Λ_5	0.86 (fixed)	1.14 (fixed)	0.88 (fixed)	1.1 (fixed)
Λ_4	0.61 (12.54)	0.79 (fixed)	0.74 (12.49)	0.63 (11.99)
Λ_3	0.99 (14.72)	1.27 (14.12)	1.01 (14.71)	1.24 (13.63)

Appendix B. Comparison of parameter ratios

In the main body of the paper, the models are compared based on their model fit. Here, we look at the parameters resulting from those models and how they compare. To compare parameters across several models, we use parameter ratio's, capturing relative importance of one parameter to another. This approach allows us also to compare the parameters of the RUM-based models to those obtained from the µRRM models (Chorus, 2012).

All models contain five taste parameters: travel time by car (Car IVT), travel time by public transport (PT IVT), Cost, public transport headway (PT headway) and mode of public transport (PT mode). From these five parameters, four ratios are obtained. In all four ratios, the “cost” parameter is the denominator and the remaining four are the respective numerators. In RUM-based models, these ratios can also be interpreted as the willingness-to-pay (WtP). While the ratios achieve a roughly similar value in RRM models as well, they cannot be interpreted as WtP (Chorus, 2012).

Fig. 5 shows the number of significantly different parameter ratios between models. The significance of the differences is determined in the same way as by Giergiczny et al. (2017). Based on the resulting p-value of the difference, we grouped them into four different groups of significance, as can be seen in the figure.

What is immediately noticeable in Fig. 5 is that for the majority of comparisons, the parameter ratios are not significantly different. Evaluating the differences between B and E models, only one parameter ratio is significantly different at the 0.1 level: the (headway/cost) ratio differs between models B1 and E1. All other parameter ratios are not significantly different even at the 0.1 level. Comparing the three B models, the only significant difference can be seen between B1 and B2, which are the (car IVT/cost) ratio at the 0.05 level and (headway/cost) ratio at the 0.01 level. Overall, the average p-value of the comparisons of the E models to B1 is 0.64, 0.72 for comparison with B2, and 0.93 when comparing parameter ratios between B3 and E models. This means that using the E models instead of B models yields a similar result in terms of respondents' taste preference, while providing further insight into the decision-making behaviour of the respondents, and in the case of model B1, also allowing the estimation with many more observations (four times as many in the case of this research).

Considering the difference in parameter ratios however, quite substantial differences between ratios of parameters from the first stage and later stages can be seen. Only five of the 12 or fewer than half of ratios are insignificantly different between those models. Significantly different are the (car IVT/cost) and (headway/cost) ratios between S1 and S2/3/4. In addition, the (mode/cost) ratio is also significantly different between models S1 and S4.

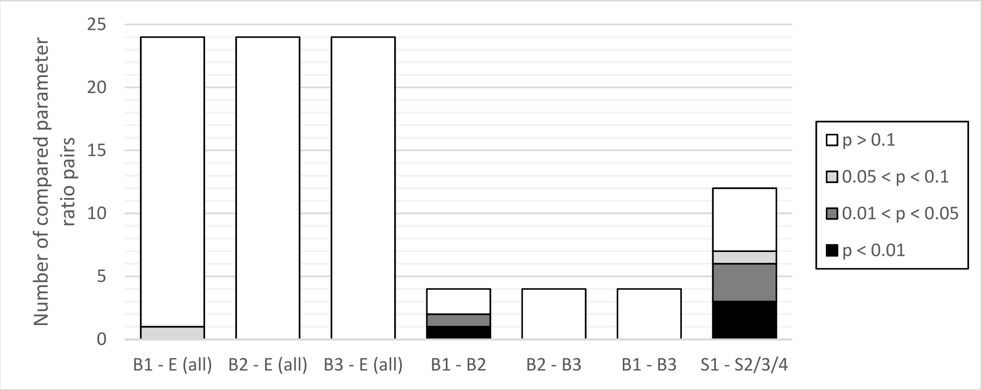


Fig. 5. Share of (in)significant parameter ratio differences between different estimated models.

References

Beach, L.R., Mitchell, T.R., 1987. Image theory: PRINCIPLES, IN decision making * goals, and plans lee Roy beach and terence R. Mitchell. Theor. Decis. 66, 201–220.

Ben-Akiva, M., Lerman, S.R., 1994. In: Discrete Choice Analysis: Theory and Application to Travel Demand, sixth ed. MIT Press, Cambridge, MA.

Ben-Akiva, M., Morikawa, T., Shiroishi, F., 1992. Analysis of the reliability of preference ranking data. J. Bus. Res. 24 (2), 149–164. https://doi.org/10.1016/0148-2963(92)90058-J.

Ben-Akiva, M., Swait, J., 1986. The Akaike likelihood ratio index. Transport. Sci. 20 (2), 133–136. https://doi.org/10.1287/trsc.20.2.133.

Bos, I.D.M., Van der Heijden, R.E.C.M., Molin, E.J.E., Timmermans, H.J.P., 2004. The choice of park and ride facilities: an analysis using a context-dependent hierarchical choice experiment. Environ. Plann. 36 (9), 1673–1686. https://doi.org/10.1068/a36138.

Chorus, C.G., 2010. A new model of random regret minimization. Eur. J. Transport Infrastruct. Res. 10 (10), 181–196.

Chorus, C.G., 2012. Random regret-based discrete choice modeling. In: SpringerBriefs in Business. https://doi.org/10.1007/978-3-642-29151-7.

Chorus, C.G., Arentze, T.A., Timmermans, H.J.P., 2008. A Random Regret-Minimization model of travel choice. Transp. Res. Part B Methodol. 42 (1), 1–18. https://doi.org/10.1016/j.trb.2007.05.004.

Chorus, C.G., Van Cranenburgh, S., Dekker, T., 2014. Random regret minimization for consumer choice modeling: assessment of empirical evidence. J. Bus. Res. 67 (11), 2428–2436. https://doi.org/10.1016/j.jbusres.2014.02.010.

Dekker, T., 2014. Indifference based value of time measures for Random Regret Minimisation models. J. Choice. Model. 12, 10–20. https://doi.org/10.1016/j.jocm.2014.09.001.

Dekker, T., Chorus, C.G., 2018. Consumer surplus for random regret minimisation models. J. Environ. Econ. Pol. 7 (3), 269–286. https://doi.org/10.1080/21606544.2018.1424039.

Dyachenko, T., Reczek, R.W., Allenby, G.M., 2014. Models of sequential evaluation in best-worst choice tasks. Market. Sci. 33 (6), 828–848. https://doi.org/10.1287/mksc.2014.0870.

Flynn, T.N., 2010. Valuing citizen and patient preferences in health: recent developments in three types of best–worst scaling. Expert Rev. Pharmacoecon. Outcomes Res. 10 (3), 259–267. https://doi.org/10.1586/erp.10.29.

Ghijben, P., Lancsar, E., Zavarsek, S., 2014. Preferences for oral anticoagulants in atrial fibrillation: a best–best discrete choice experiment. Pharmacoeconomics 32 (11), 1115–1127. https://doi.org/10.1007/s40273-014-0188-0.

Giergiczny, M., Dekker, T., Hess, S., Chintakayala, P.K., 2017. Testing the stability of utility parameters in repeated best, repeated best-worst and one-off best-worst studies. Eur. J. Transport Infrastruct. Res. 17 (4) https://doi.org/10.18757/ejtir.2017.17.4.3209.

Guevara, C.A., Chorus, C.G., Ben-Akiva, M.E., 2016. Sampling of alternatives in random regret minimization models. Transport. Sci. 50 (1), 306–321. https://doi.org/10.1287/trsc.2014.0573.

Hauser, J.R., 2008. Testing the accuracy, usefulness, and significance of probabilistic choice models: an information-theoretic approach. Oper. Res. 26 (3), 406–421. https://doi.org/10.1287/opre.26.3.406.

Lancsar, E., Fiebig, D.G., Hole, A.R., 2017. Discrete choice experiments: a guide to model specification, estimation and software. Pharmacoeconomics 35 (7), 697–716. https://doi.org/10.1007/s40273-017-0506-4.

Lancsar, E., Louviere, J., Donaldson, C., Currie, G., Burgess, L., 2013. Best worst discrete choice experiments in health: methods and an application. Soc. Sci. Med. 76 (1), 74–82. https://doi.org/10.1016/j.socscimed.2012.10.007.

Lancsar, E., Louviere, J.J., 2008. Estimating Individual Level Discrete Choice Models and Welfare Measures Using Best-Worst Choice Experiments and Sequential Best-Worst MNL. Sydney.

Louviere, J.J., Street, D., Burgess, L., Wasi, N., Islam, T., Marley, A.A.J., 2008. Modeling the choices of individual decision-makers by combining efficient choice experiment designs with extra preference information. J. Choice. Model. 1 (1), 128–164. https://doi.org/10.1016/S1755-5345(13)70025-3.

Marley, A.A.J., Louviere, J.J., 2005. Some probabilistic models of best, worst, and best-worst choices. J. Math. Psychol. 49 (6), 464–480. https://doi.org/10.1016/j.jmp.2005.05.003.

Meloy, M.G., Russo, J.E., 2004. Binary choice under instructions to select versus reject. Organ. Behav. Hum. Decis. Process. 93 (2), 114–128. https://doi.org/10.1016/j.obhdp.2003.12.002.

Ordóñez, L.D., Benson, L., Beach, L.R., 1999. Testing the compatibility test: how instructions, accountability, and anticipated regret affect prechoice screening of options. Organ. Behav. Hum. Decis. Process. 78 (1), 63–80. https://doi.org/10.1006/obhd.1999.2823.

Stone, M., 1979. Comments on model selection criteria of Akaike and Schwarz. J. Roy. Stat. Soc. B 276–278.

Train, K.E., 2009. Discrete choice methods with simulation. In: Discrete Choice Methods with Simulation, vol. 9780521816. https://doi.org/10.1017/CBO9780511753930.

Tversky, A., 1972. Elimination by aspects: a theory of choice. Psychol. Rev. 79 (4), 281–299. https://doi.org/10.1037/h0032955.

van Cranenburgh, S., Collins, A.T., 2019. New software tools for creating stated choice experimental designs efficient for regret minimisation and utility maximisation decision rules. J. Choice. Model. 31, 104–123. https://doi.org/10.1016/j.jocm.2019.04.002.

Van Cranenburgh, S., Guevara, C.A., Chorus, C.G., 2015a. New insights on random regret minimization models. Transport. Res. Pol. Pract. 74, 91–109. https://doi.org/10.1016/j.tra.2015.01.008.

Van Cranenburgh, S., Prato, C.G., Chorus, C., 2015b. Accounting for Variation in Choice Set Size in Random Regret Minimization Models. Retrieved from. http://pure.tudelft.nl/ws/files/7828025/Accounting_for_variation_in_choice_set_size_in_RRM_models_29072015.pdf.

Van Cranenburgh, S., Rose, J.M., Chorus, C.G., 2018. On the robustness of efficient experimental designs towards the underlying decision rule. Transport. Res. Pol. Pract. 109, 50–64. https://doi.org/10.1016/j.tra.2018.01.001. March 2017.

Walker, J.L., Wang, Y., Thorhauge, M., Ben-Akiva, M., 2018. D-efficient or deficient? A robustness analysis of stated choice experimental designs. Theor. Decis. 84 (2), 215–238. https://doi.org/10.1007/s11238-017-9647-3.

Zeelenberg, M., Pieters, R., 2007. A theory of regret regulation 1.0. J. Consum. Psychol. 17 (1), 3–18. https://doi.org/10.1207/s15327663jcp1701_3.