On the robustness of efficient experimental designs towards the underlying decision rule

Sander van Cranenburgh[a], John M. Rose[b], Caspar G. Chorus[a]

[a]*Transport and Logistics Group, Delft University of Technology*

[b]*Institute of Choice, University of South Australia*

1 Introduction

Stated Choice (SC) experiments are widely used to acquire understanding on travel behaviour (Louviere et al. 2000). Nowadays, SC experiments are increasingly being based on so-called "efficient designs". Efficient designs aim to generate stated choice tasks that maximize the collected information in the data, yielding more reliable parameter estimates with an equal, or lower number of observations than traditional orthogonal designs (Rose and Bliemer 2009; Kessels et al. 2011). While earlier research efforts on efficient experimental design mainly focussed on extending the design theory to encompass more advanced models of choice, such as Nested Logit and (Panel) Mixed Logit models (e.g. Yu et al. 2009; Bliemer and Rose 2010), recent efforts are shifting towards understanding the robustness of the modelling outcomes towards the experimental design (Ferrini and Scarpa 2007; Bliemer and Rose 2011; Rose and Bliemer 2014).

The literature exploring the robustness of the modelling outcomes with respect to the experimental design has predominantly focussed on misspecification in terms of the parameter priors and the way in which (correlations between) error terms are modelled. Yet, despite the compelling evidence from the fields of transport showing that travellers use a wide range of decision rules when making choices (Hess et al. 2012; Boeri et al. 2014), robustness issues concerning potential misspecification of the presumed decision rule has attracted only very limited attention within the literature (see Rose and Bliemer 2013a). In fact, to the authors' knowledge, research on experimental designs has exclusively been based on the (often implicit) assumption that decision-makers make choices based on (linear-additive) Random Utility Maximization (RUM). As a consequence, it is currently unclear what is the influence of different assumptions regarding the decision rules on the statistical efficiency of the design (e.g. does a misspecification of the decision rule result in different, and perhaps highly suboptimal, designs?).

This paper aims to fill these knowledge gaps. To do so, we construct efficient designs based on a non-RUM model – in casu: a Random Regret Minimization (RRM) model (Chorus 2010) – and assess the effects of decision rule misspecification analytically as well as empirically. We use an RRM model for our analyses because RRM models are among the more popular non-RUM models. Moreover, the specific RRM model we use: the P-RRM model (Van Cranenburgh et al. 2015), is equally parsimonious as the canonical linear-additive RUM model, and has very convenient mathematical properties for constructing efficient designs. First, we investigate analytically the influence of the assumption regarding the underlying decision rule on the statistical efficiency. Specifically, we consider two cases: (1) the case in which the experimental designs are optimized for linear-additive RUM while the true Data Generating Process (DGP) is P-RRM, and (2) the case in which the experimental efficient designs are optimized for P-RRM while the true DGP is linear-additive RUM. After that, we use empirical data to investigate the impacts on estimation outcomes. Specifically, we analyse and compare the results of two SC experiments which differ only in terms of the decision rule that is being used to construct the efficient experimental design.

The methodological contributions of this paper to the experimental design literature are twofold. Firstly, we show that for the P-RRM model (Van Cranenburgh et al. 2015) efficient designs can relatively easily be constructed. Because the P-RRM model has a piecewise linear form, the Asymptotic Variance Covariance matrix (AVC) – which is needed to construct efficient designs – can be determined *analytically*, just like for the linear-additive RUM-MNL models (in contrast, to generate efficient designs for other types of RRM models *simulation* of the AVC is required). Secondly, we extend the experimental literature by developing new insights on the effects of misspecification of the assumed underlying decision rule on the statistical efficiency. Finally, the substantive contribution of this paper is that we develop new empirical insights on the robustness of estimation outcomes with respect to the assumed underlying decision rule used to create the efficient design.

The remainder of this paper is structured as follows. Section 2 discusses efficient design theory. Section 4 analytically explores the influence of the assumption regarding the underlying decision rule on the statistical efficiency. Section 5 presents an empirical case study. Section 6 summarizes our findings and reports our conclusions.

2 Efficient designs

2.1 Theory

Designing an SC experiment involves making a number of decisions, such as how many choice tasks are presented to each respondent, how many alternatives per choice task, and what are the attributes considered, and what are their levels. After having determined this broad design set up, the next step is to select a design strategy. Two designs strategy are predominantly used in the literature: 1) based in the principle of orthogonality, and 2) based on statistical efficiency. Orthogonal designs occur when each attribute level combination over the design occurs an equal number of times, and has sometime been mistaken to be any design that minimises the correlation between the attribute levels in the choice tasks, while statistically efficient designs aim to generate designs that maximize the collected information in the data. The premise underlying statiscal designs is that by doing so more reliable parameter estimates can be attained with an equal, or lower number of observations than when orthogonal designs would have been used (Rose and Bliemer 2009). In other words, efficient designs aim to minimize the (asymptotic) standard errors, or alternatively, maximize the t-ratios, of the parameters estimates of the model under consideration. In effect, ultimately both design paradigms may be classed as efficient designs however, as orthogonal designs can be shown to be efficient (optimal) designs under certain assumptions (see Rose and Bliemer 2013b).

Statistical efficiency can be measured in several ways such as A-efficiency, C-efficiency, D-efficiency, and S-efficiency. Regardless of the exact measure that is being used, they exploit the Asymptotic Variance Covariance (AVC) matrix of the parameters estimates to determine the efficiency. The AVC matrix is a function of the design itself, the (prior) parameter estimates, and the model specification, and is equal to the negative inverse of the expected Fisher Information matrix, see equation 1. In turn, the expected Fisher Information matrix can be computed by taking the second order derivatives of the Log-Likelihood function w.r.t. the model's parameters.

$$AVC = -\left(E\left[I\left(X,Y,\beta\right)\right]\right)^{-1} = -\left[\frac{\partial^2 LL\left(X,Y,\beta\right)}{\partial^2 \beta}\right]^{-1} \tag{1}$$

The Log-Likelihood function for a discrete choice model is given by equation 2. When all respondents face the same choice tasks – as is usually the case in SC experiments – only one design needs to be evaluated. Hence, the first summation can be dropped, yielding equation 3.

$$LL\left(\beta\right) = \sum_{n=1}^{N}\sum_{s=1}^{S}\sum_{j=1}^{J} y_{nsj} \ln\left(P_{nsj}\right) \tag{2}$$

$$LL\left(\beta\right) = \sum_{s=1}^{S}\sum_{j=1}^{J} y_{nsj} \ln\left(P_{nsj}\right) \tag{3}$$

In the context of this research we focus on one measure for statistical efficiency: the D-error statistic (D-efficiency). The D-error statistic is the most commonly used measure of efficiency in experimental design practice. The D-error statistic is calculated by taking the determinant of the Asymptotic Variance Covariance (AVC) matrix. Accordingly, in the context of this study the optimal efficient design $D_{optimal}$ is that design that attains the lowest D-error, see equation 4. Finally, it is important to note that technically we minimize the *expected* determinant of the AVC, as prior to estimation neither the exact model specification nor the model's parameters are known.

$$D_{optimal} = \arg\min_X \left(\det\left[AVC\right]\right) \tag{4}$$

2.2 Efficient designs for Random Regret minimization models

This section derives the AVC matrix for the P-RRM model, which is needed to construct efficient designs based on the P-RRM model. We show that for the P-RRM model the AVC matrix is independent of the choices that are eventually being made by the respondents, and therefore can relatively straightforwardly be derived analytically, just like for linear-additive RUM Multinomial Logit (MNL) models (Huber and Zwerina 1996), linear-additive RUM Nested Logit models (Bliemer et al. 2009), and the linear-additive RUM Mixed MNL models (see e.g. Sandor and Wedel 2005). This is an important feature, considering the fact that in order to find an efficient design typically many thousands candidate designs need to be evaluated.

2.2.1 The P-RRM model

RRM models are based on the premise that, when choosing, the decision maker *n* minimizes regret. Regret is experienced when a competitor alternative *j* outperforms the considered alternative *i* with regard to attribute *m*. The overall regret of an alternative is conceived to be the sum of all the pairwise regrets that are associated with bilaterally comparing the considered alternative with the other alternatives in the choice set in terms of each of the attributes. The general form of RRM models is given in equation **5**, where RR_{in} denotes the random regret experienced by decision maker *n* considering alternative *i*, R_{in} denotes the observed part of regret, and ε_{in} denotes the unobserved part of regret. In the core of RRM models[1] is the so-called attribute level regret function $r_{ijmn} = f\left(\beta_m, x_{jmn} - x_{imn}\right)$. This function maps the difference between the levels of attributes *m* of the competitor alternatives *j* and the considered alternative *i* onto regret.

$RR_{in} = R_{in} + \varepsilon_{in} \; where \; R_{in} = \sum_{j \neq i} \sum_{m} r_{ijmn}$	5

In the P-RRM model the attribute level regret function takes the form of equation **6**, and is depicted in Figure 1. This model is a special case of the highly flexible μRRM model[2] (Van Cranenburgh et al. 2015; and see Van Cranenburgh and Prato 2016 for an overview of RRM models), and has a cornerstone meaning within the RRM modelling paradigm. Since it postulates no rejoice (the opposite of regret), it yields the strongest regret minimization behaviour (most regret aversion) possible within the RRM framework. Furthermore, it is important to mention that the model is equally parsimonious as the canonical linear-additive RUM model.

$r_{ijmn} = \max\left(0, \beta_m \left[x_{jmn} - x_{imn}\right]\right)$	6

[1] One exception to this form is the RRM specification proposed in Chorus et al. (2008).
[2] The μRRM model also nests the linear-additive RUM model.

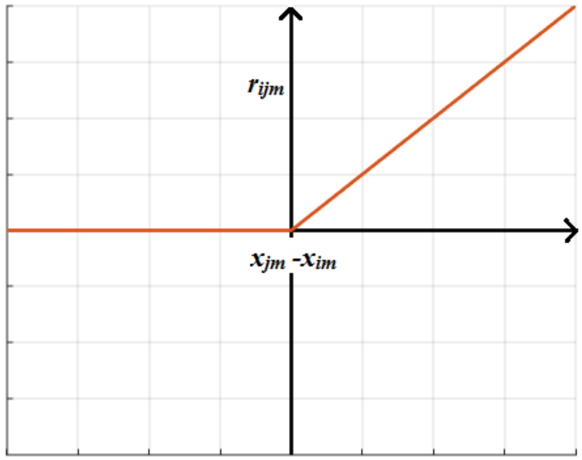


Figure 1: Attribute level regret function of the P-RRM model

At first sight, the max operator in the attribute level regret function (equation **6**) may seem undesirable from a computational perspective. However, as shown by Van Cranenburgh et al. (2015), when the signs of the taste parameters are a priori known to the researcher – as is usually the case in a transport setting – the β_m's can be placed outside the max operator, see equation **7**, where β_m^+ denotes positive taste parameters, and β_m^- denotes negative taste parameters.

$$\begin{aligned} R_{in} &= \sum_m \sum_{j \neq i} \max\left(0, \beta_m \left[x_{jmn} - x_{imn}\right]\right) \\ &= \sum_{m^+} \beta_m^+ \sum_{j \neq i} \max\left(0, \left[x_{jmn} - x_{imn}\right]\right) + \sum_{m^-} \beta_m^- \sum_{j \neq i} \min\left(0, \left[x_{jmn} - x_{imn}\right]\right) \end{aligned} \tag{7}$$

Conveniently, the terms $\sum_{j \neq i} \max\left(0, \left[x_{jmn} - x_{imn}\right]\right)$ and $\sum_{j \neq i} \min\left(0, \left[x_{jmn} - x_{imn}\right]\right)$ simply constitute data operations, implying that they can be computed *prior to the estimation*. Therefore, the resulting functional form of the P-RRM model has a linear-additive form (equation **8**).

$$R_{in}^{P-RRM} = \sum_m \beta_m x_{imn}^{P-RRM} \quad where\ x_{imn}^{P-RRM} = \begin{cases} \sum_{j \neq i} \max\left(0, x_{jmn} - x_{imn}\right) & if\ \beta_m > 0 \\ \sum_{j \neq i} \min\left(0, x_{jmn} - x_{imn}\right) & if\ \beta_m < 0 \end{cases} \tag{8}$$

An inherent difficulty in efficient design is setting the priors. This is the case in general, but in the context of this study, in which we explore the relative efficiencies of RUM and P-RRM designs. To set the priors ‘equal’ in the RUM and P-RRM case we apply a ‘choice set size correction’ factor, see equation **9** where $\tilde{R}_{in}$ denotes the choice set size corrected regret, Γ denotes an arbitrary constant, and J_n denotes the choice set size in observation n. This correction factor is developed by see Van Cranenburgh et al. (working paper) to be able to estimate RRM models on a data set in which the number of alternatives that are available to a decision maker vary across choice observations. However, also in this context the correction factor comes in handy as it takes out the the effect of the choice set size on the sizes of RRM model parameters. More specifically, we use the following parametrization: $\Gamma = 2$. Using this parametrization, we ensure that the RUM and P-RRM model parameters are fully equivalent to one another in the binary choice context.

$$\tilde{R}_{in} = \frac{\Gamma}{J_n} \cdot R_{in} \qquad (9)$$

Accordingly, choice set size corrected regrets in the P-RRM model are given by equation **10**.

$$\tilde{R}_{in}^{P-RRM} = \sum_m \beta_m \tilde{x}_{imn}^{P-RRM} \quad where\ \tilde{x}_{imn}^{P-RRM} = \begin{cases} \dfrac{\Gamma}{J_n} \sum_{j \neq i} \max\left(0, x_{jmn} - x_{imn}\right)\ if\ \beta_m > 0 \\ \dfrac{\Gamma}{J_n} \sum_{j \neq i} \min\left(0, x_{jmn} - x_{imn}\right)\ if\ \beta_m < 0 \end{cases} \qquad (10)$$

Finally, inspired by the RUM modelling practice, commonly the negative of the error term in RRM models is assumed to be i.i.d. type I Extreme Value distributed with a variance of $\pi^2/6$, resulting in the well-known and convenient closed-form logit formula for the choice probabilities given by equation **11**. Note that since RRM models assume minimization of regret, a minus sign is placed in front of the observed regret.

$$P_{in} = \frac{e^{-\tilde{R}_{in}}}{\sum_J e^{-\tilde{R}_{jn}}} \qquad (11)$$

2.2.2 *Deriving the AVC matrix for the P-RRM MNL model*

Next, we derive the AVC for the P-RRM-MNL model. To do so, we first derive the first order derivative of the Log-Likelihood function with respect to the model parameter β. After that, we derive the second order derivative of the Log-Likelihood function with respect to β, which is in turn needed to compute the AVC matrix. Note that it is only required to derive the derivatives w.r.t. *generic* RRM parameters: due to the binary nature of attributes associated with non-generic parameters (such as ACSs and dummies), RUM and RRM treatment of these attributes is mathematically equivalent (see e.g. Chorus 2012a).

<u>First order derivative</u>

The simplest way to find the first order derivatives of the Log-Likelihood function is to take advantage of the results for the linear-additive RUM-MNL model, and apply the chain rule. The first order derivative of the Log-Likelihood of the linear-additive RUM–MNL model is given in equation **12**:

$$\frac{\partial LL^{RUM}}{\partial \beta_m} = \frac{\partial LL}{\partial V_{js}} \cdot \frac{\partial V_{js}}{\partial \beta_m} = \sum_{s=1}^{S} \sum_{j=1}^{J} \left(y_{js} - P_{js} \right) x_{jms} \tag{12}$$

Substitution of $V_{js} = -R_{js}$ and applying the chain rule[3], directly yields the first order derivative of the Log-Likelihood function of the P-RRM MNL model with respect to generic parameter β_k (equation **13**):

$$\begin{aligned} \frac{\partial LL^{P-RRM}}{\partial \beta_m} &= \frac{\partial LL}{\partial R_{js}} \cdot \frac{\partial R_{js}}{\partial V_{js}} \cdot \frac{\partial V_{js}}{\partial \beta_m} = \frac{\partial LL}{\partial V_{js}} \cdot \frac{\partial V_{js}}{\partial R_{js}} \cdot \frac{\partial R_{js}}{\partial \beta_m} \\ &= -\sum_{s=1}^{S} \sum_{j=1}^{J} \left(y_{js} - P_{js} \right) \tilde{x}_{jms}^{P-RRM} \end{aligned} \tag{13}$$

Hence, the only difference as compared to the first order derivative of the linear-additive RUM MNL is the appearance of a minus sign. This is in line with intuition given the facts that (1) regret

[3] Note that: $\frac{\partial V_{js}}{\partial \beta_k} = x_{jks}$

can be conceived as negative utility, and (2) under the P-RRM model the regrets are linear-additive.

Second order derivative

The second order derivative is somewhat more cumbersome to derive than the first order derivative. For reasons of brevity the full derivation can be found in Appendix X. The second order derivative of the Log-Likelihood function of the P-RRM-MNL model w.r.t. its (generic) taste parameters is also very similar to its linear-additive RUM-MNL counterpart. In fact, the only difference is the replacement of the attribute levels x_{jms}, by the P-RRM model's attribute levels $\tilde{x}_{jms}^{P-RRM}$, see equation **14**.

$$\frac{\partial^2 LL}{\partial\beta_{m_1}\partial\beta_{m_2}} = -\sum_{s=1}^{S}\sum_{j=1}^{J}\tilde{x}_{jm_1s}^{P-RRM} P_{js}\left(\tilde{x}_{jm_2s}^{P-RRM} - \sum_{i=1}^{J} x_{im_2s}^{P-RRM} P_{is}\right) \tag{14}$$

3 Exploring the robustness of efficient designs towards misspecification of the underlying decision rule

This section explores the robustness of efficient designs towards misspecification of the underlying decision rule. In particular, we analytically assess how efficient linear-additive RUM optimal efficient designs are when the true DGP is P-RRM, and vice versa, how efficient P-RRM optimal efficient designs are when the true DGP is linear-additive RUM. To do so, we use the following two measures of robustness.

- *Relative sub optimality.* Relative sub optimality is defined as the percentage of the total number of designs that would perform better than the optimal efficient based on the incorrectly specified decision rule.

- *Efficiency loss.* This is the decrease in the D-efficiency due to using a misspecified decision rule, relative to the efficiency that could have been obtained when the correct design rule would have been used. Equation **15** gives the loss function, where $\tilde{D}$ is the optimal D-error statistic obtained using the *incorrect* decision rule, and $\bar{D}$ is the optimal D-error statistic obtained using the *correctly* specified decision rule. Note that this implies that we need to know the optimal D-error. While this is typically not the case, in this study we focus on a relatively small design tasks, and are able to establish the optimal D-error.

$L = 1 - \frac{\bar{D}}{\tilde{D}}$	15

To explore the robustness of efficient designs towards misspecification of the underlying decision rule both measure are insightful. The relative sub optimality is an intuitive measure, as it shows the percentage of designs that would perform better than the optimal efficient based on the incorrectly specified decision rule. However, while intuitive, to fully grasp the effect of misspecification of the decision rule on the statistical efficiency also the loss in efficiency needs to be considered. After all, finding for instance a relative sub optimality of 50%, does not necessarily imply a highly statistically inefficient design was obtained using the incorrect decision rule as the distribution of the D efficiencies across all designs can be such that many designs have efficiencies that are close to the optimal efficient design.

3.1 Experimental design tasks

We consider two relatively small experimental design tasks. By doing so we are able to determine the optimal efficient designs i.e. the most efficient design of all possible designs. This is crucial in this context, as we thereby avoid the possibility of presenting misguided results due to the possibility that the search algorithm (which otherwise would have been used) may not have been able to find the optimal solution. In fact, in the context of large solution spaces search algorithms are only able to assess a very tiny part of the solution space. Therefore, it is likely that the search algorithm is not able to find the truly optimal efficient solution. Moreover, search algorithms may get ‘stuck’ in local optimal, although they typically frequently reseed to avoid this problem (Palhazi Cuervo et al. 2014).

More specifically, Table 1 presents the experimental design tasks. The two design tasks both comprise of three alternatives each alternative defined by two generic attributes (e.g. cost and time). We consider three alternatives as this is the minimum number of alternatives that is required to be able distinguish between RUM behaviour and RRM behaviour. After all, a distinguishing property of RRM models as compared to RUM models is that they are able to capture context effects, such as the compromise effect (Chorus 2012b). Such context effects only manifest in the presence of three or more alternatives. Furthermore, the alternatives comprise of two attributes, as this is the minimum number of attributes needed to analyse trade-offs in discrete choice models. Design tasks 1 and 2 differ from one another in terms of the number of attribute levels, and the number of choice tasks per design. In design task 1 four attribute levels are used, and four choice tasks per respondent, while in design task 2 five attribute levels and three choice tasks per respondent[4].

Table 1: Design tasks

	Design task 1	**Design task 2**
Number of alternatives	3	3
Number of generic attributes	2	2
Number of attribute levels per attribute	4	5
Number of choice tasks per respondent	4	3
Attribute levels	1,2,3,4	1,2,3,4,5
Unique alternatives	4^2=16	5^2=25
Number of choice tasks (i.e. the full factorial design: L^{MA})	$4^{3\cdot 2}=4,096$	$5^{3\cdot 2}=15,625$
Number of unique choice tasks, without duplicate	560	2,300
Number of designs	$\binom{560}{4} \approx 4 \times 10^9$	$\binom{2,300}{3} \approx 2 \times 10^9$
Number of unique choice tasks, without duplicate or dominant alternatives	16	100
Number of designs, without duplicate or dominant alternatives	$\binom{16}{4} = 1,820$	$\binom{100}{3}$ =161,700

In total 16 and 25 unique alternatives can be constructed for design task 1 and 2, respectively. As a consequence, there are respectively 4,096 and 15,625 of possible choice tasks. However, those numbers can considerably be reduced, as the order in which alternatives are placed in the choice task is irrelevant in terms of statistical efficiency and as duplication of alternatives in choice tasks

[4] Note that we are not concerned with attribute level balance.

is undisirable. Removing these choice tasks leaves us with respectively 560 and 2,300 unique choice tasks. Based on these choice tasks, we can construct respectively 4 billion and 2 billion possible designs for design tasks 1 and 2.

We further limit the solution space by removing choice tasks containing one or more dominant alternatives.[5] The first and most important reason for doing so is that dominant alternatives are know to bias the parameters estimates (Huber et al. 1982; Bliemer and Rose 2011; Bliemer et al. 2016). Conventional discrete choice models are simply not well-equipted to deal with dominant alternatives. Therefore, it is common practice to remove choice tasks containing dominant alternatives[6] (Hensher et al. 1988; Kouwenhoven et al. 2014), although sometimes a dominant choice task is added to see whether respondents understood the choice task (see e.g. Bradley and Daly 1994). A second reason relates to the fact that respondents may get agitated by choice tasks containing dominant alternatives, especially in the case of relatively simple SC experiments which comprise is just two or three attributes, such as ours. A final, pragmatic reason for removing choice tasks containing dominant alternatives is that it vastly reduces the solution space. In fact, removing those choice tasks containing dominant alternatives reduces number of unique choice tasks to a mere 16 in design task 1 and to 100 choice tasks in design task 2. In turn, are just 1,820 and 161,700 designs for design task 1 and task 2, respectively. This enables us to quickly identify the optimal design, and to explore the robustness of efficient designs towards the decision rule using series of parametrizations.

Finally, we use perfect priors. That is, the priors used to construct the efficient designs are equal to the parameters used in the DGP. The RUM and P-RRM parameters are set to the same values. While there is no strong theoretical ground for presumption (other than in binary choice situations, see Chorus 2010), a substantial body of empirical evidence suggests that – after correcting for the choice set size (see section 2.2.1) – this is a reasonable assumption.

3.2 Analytical results

Figure 2 and Figure 3 show the sub optimality for respectively design tasks 1 and 2, as a function of the model parameters β_1 and β_2; The left hand plots show the sub optimality of P-RRM models estimated on designs that are optimal efficient for linear-additive RUM; the right hand plots show

[5] An alternative A is said to dominante alternative B if alternative A outperforms alternative B in terms of all attributes. Hence, we presume the signs of the two taste parameters are known in advance.
[6] In contrast to our approach, this is commonly done after generating the experimental designs

the sub optimality of linear-additive RUM models estimated on designs that are optimal efficient for P-RRM models. Figure 4 and Figure 5 show the loss in efficiency for respectively design tasks 1 and 2, as a function of the model parameters β_1 and β_2. Similar as in Figure 2 and Figure 3, the left hand plots show the loss in efficiency of P-RRM models estimated on designs that are optimal efficient for linear-additive RUM; the right hand plots show the sub optimality of linear-additive RUM models estimated on designs that are optimal efficient for P-RRM models. To explore the effect of the size and the ratios of the parameters, the parameter space ranges from 0 to 3, for both taste parameters in all plots.

Based on Figure 2 to Figure 5 a number of observations can be made. Firstly, there are parameter spaces where linear-additive RUM optimal efficient designs are highly inefficient when the true DGP is P-RRM, and there are parameter spaces where P-RRM optimal efficient designs are highly inefficient when the true DGP is RUM. In fact, Figure 2 and Figure 3 show that a design optimal efficient for linear-additive RUM can be among the worst of the set of feasible designs for in case the true DGP is P-RRM (e.g. for $\beta_1 = \beta_2 = 2$). Likewise, a optimal efficient design for P-RRM can be among the worst of the set of feasible designs for in case the true DGP is RUM (e.g. for $\beta_1 = \beta_2 = 2.5$).

Secondly, in both design tasks we see that the efficiency of a misspecified design (in terms of the underlying decision rule) is highly sensitive towards the exact combination of the parameter estimates. The relative sub optimality and the loss in efficiency can jump very abruptly when the parameters change only slightly. Furthermore, note that Figure 2 to Figure 5 are symmetric along the line y = x. This is due to the fact that the ranges of the attribute levels and parameters are equal for attributes.

All in all, these analytic results suggest that efficient designs can be highly non-robust towards decision rule misspecification. In other words, efficient experimental designs can be statistically highly inefficient when the DGP used to construct the design does not correspond with the DGP used by the decision maker.

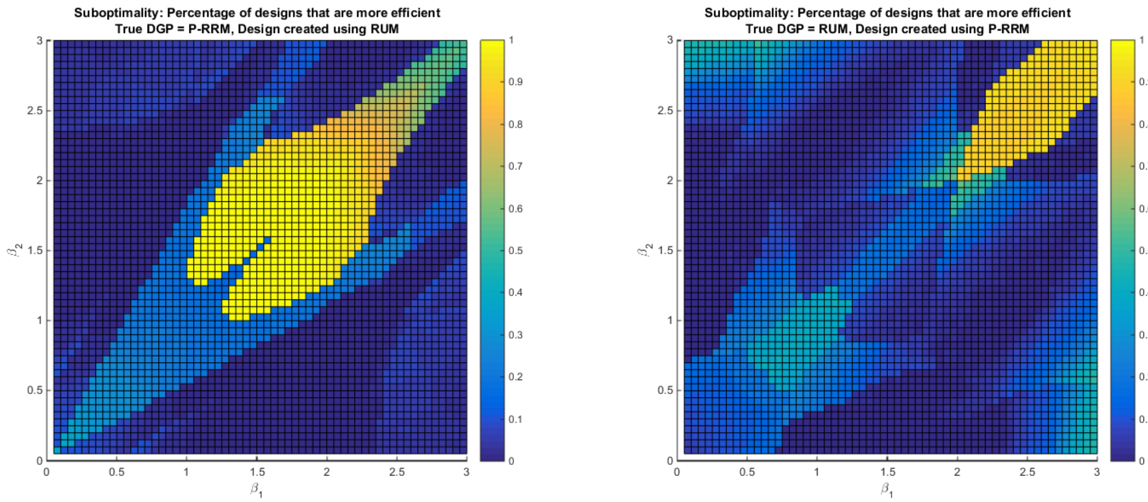


Figure 2 Relative sub optimality design task 1

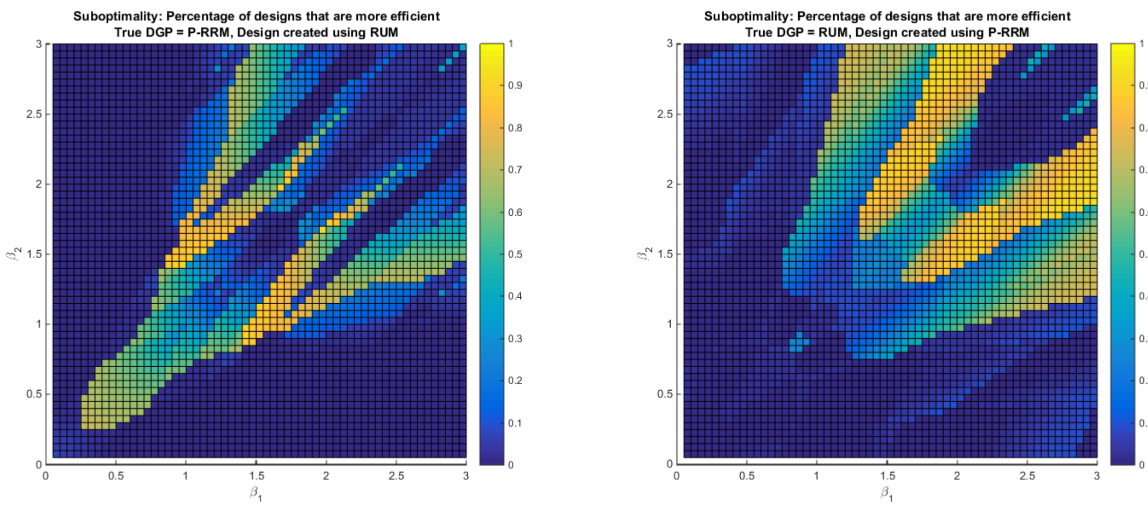


Figure 3 Relative sub optimality design task 2

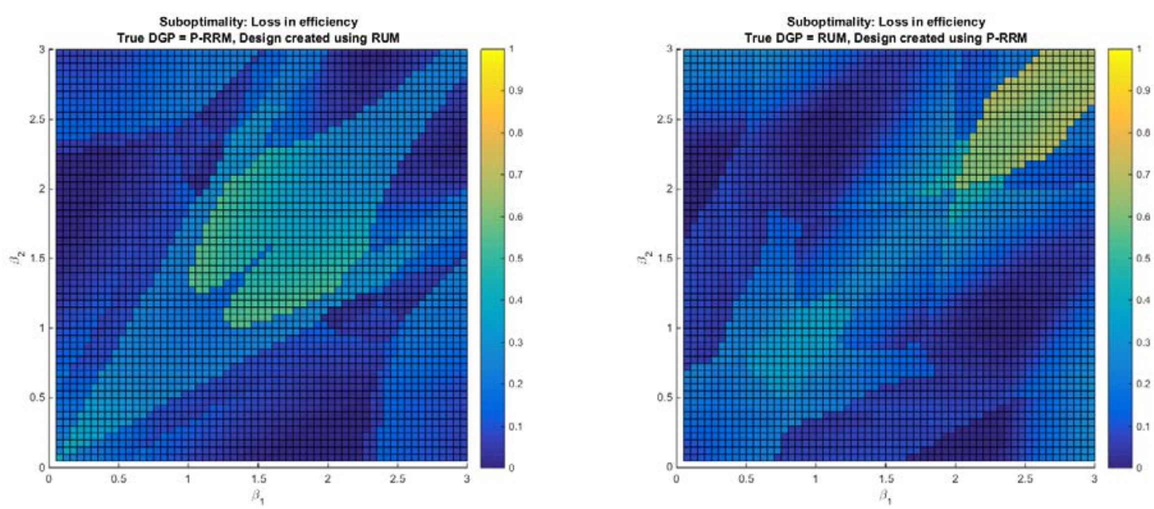


Figure 4 Loss in efficiency design task 1

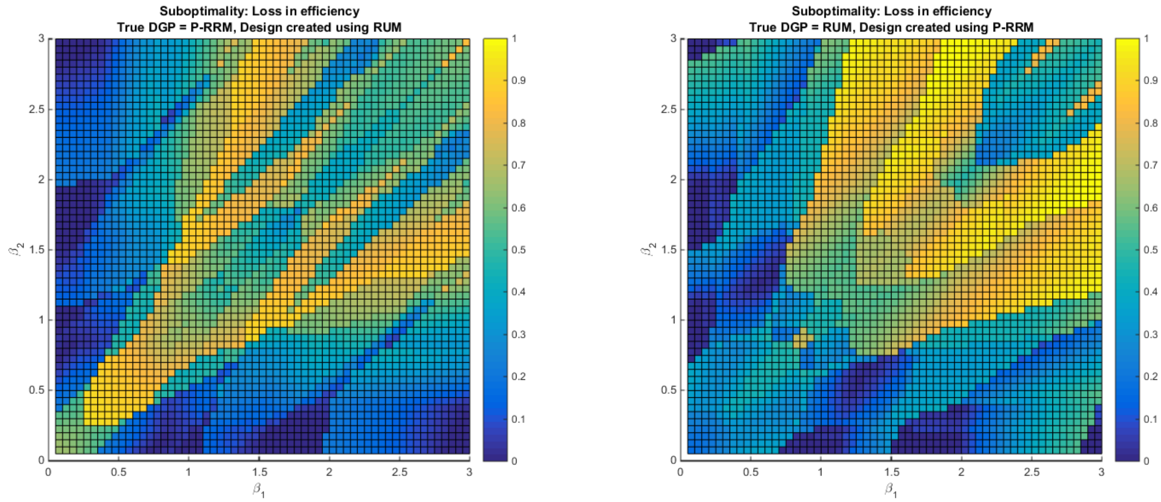


Figure 5 Loss in efficiency design task 2

4 Case study: Value-of-Travel time

To empirically investigate the robustness of efficient experimental designs towards misspecification of the underlying decision rule, we conducted a case study: a Value-of-Travel Time (VoT) SC experiment. SC experiments are widely used to infer the VoT, which, in turn, plays a crucial role in transport infrastructure evaluations (Hess et al. 2005; Abrantes and Wardman 2009; Börjesson and Eliasson 2014; Kouwenhoven et al. 2014; Ojeda-Cabral et al. 2016a). In addition, VoT SC experimental designs are typically relatively small designs in terms of the number of alternatives and attributes. This makes VoT SC designs particularly suitable in this context and congruent with our analytical analyses in section 4. Lastly, relatively recent VoT priors are available which we can to construct the efficient designs (Kouwenhoven et al. 2014).

4.1 The experimental design

Table 2 shows the SC design task, which is by and large in line with current European practice. To infer the VoT we present respondents route choices. Each route choice task consists of 3 unlabelled alternatives. Although, the majority of the European VoT SC experiments involve binary route choice tasks, we divert from this practice because regret minimization and utility maximization cannot be distinguished from one another based on binary choice data. Furthermore, we use just two attributes: travel time and travel cost. Attribute levels were selected as follows: the range of the travel times was chosen such that they are in consonance with the range of the travel times presented in previous European SC-experiments, which is usually in the order of of 10 to 15 minutes. The minimum travel time was set at 23 minutes, and the maximum at 35 minutes, with an equally spaced 4 minutes interval. By using 4 minutes intervals we avoid unresolved issues relating to the valuation of small travel time savings (Welch and Williams 1997; Daly et al. 2014). Accordingly, there are 16 unique alternatives. In turn, there are 560 unique choice tasks, without duplicate alternatives. Finally, we remove the choice tasks with dominant alternatives (for reasons discussed in section 3.1). As a consequence, there are in total 16 choice tasks to design the experiments. Each design consists of 4 choice tasks, implying a total of 1,820 possible designs.

Table 2: Experimental design task VoT

Number of alternatives	3
Number of generic attributes	2
Number of attribute levels per attribute	4
Number of choice tasks per respondent	4
Attribute levels Time [minutes]	23, 27, 31, 35
Attribute levels Cost [euro]	3, 4, 5, 6
Unique alternatives	$4^2=16$
Number of unique choice tasks, without duplicate alternatives	560
Number of designs	$\binom{560}{4} \approx 4 \times 10^9$
Number of unique choice tasks, without duplicate or dominant alternatives	16
Number of designs, without duplicate or dominant alternatives	$\binom{16}{4} = 1{,}820$

To construct the VoT efficient designs, we need priors for β_{Time} and β_{Cost}. In the most recent Dutch VoT study (Kouwenhoven et al. 2014), a VoT of €9 per hour (€0.15/minute) was found for car drivers in the commute. However, as shown by Figure 2 to Figure 5, not only the ratio of the parameters, but also the scale (thus the amount of observed versus the amount of unobserved utility) matters for the experimental design. For linear-additive RUM-MNL any combination β_{Time} and β_{Cost} that satisfies $\beta_{Time}/\beta_{Cost} = 0.15$, yields a VoT of €9.[7] In other words, we need to set

[7] In the context of RRM VoT is not univocally defined, see Dekker (2014) for more details.

one of the βs so that other β is determined. Based on a pretest under graduate students, we decided to set β_{Time} to -0.15, implying β_{Cost} = -1.00, both for the RUM and P-RRM designs.

To find the most efficient RUM and P-RRM designs we computed the D-error statistics for all 1,820 possible designs, given our priors. The designs with the lowest D-error are selected. As we focus exclusively on statistical efficiency, no attention was paid to attribute level balance, utility balance, and so on. Table 3 shows the obtained most efficient RUM (left) and P-RRM (right) experimental designs. These designs are used in our SC experiments. In line with the findings in section 3, the D-error statistics show that designs optimized for RUM models can be relatively inefficient for P-RRM models, and vice versa. The loss of efficiency in case the true DGP is P-RRM and this RUM design is used is $L = 1 - 0.0110/0.0194 = 0.57$; the loss of efficiency in case the true DGP is RUM and this P-RRM design is used is $L = 1 - 0.0178/0.0317 = 0.44$. Furthermore, we see that the designs have one choice task in common (choice task number 2).

Table 3: Experimental designs

	RUM-MNL DESIGN						P-RRM-MNL DESIGN					
D $_{RUM}$	0.0178						0.0317					
D $_{RRM}$	0.0194						0.0110					
Choice	**Route A**		**Route B**		**Route C**		**Route A**		**Route B**		**Route C**	
task	**TT**	**TC**	**TT**	**TC**	**TT**	**TC**	**TT**	**TC**	**TT**	**TC**	**TT**	**TC**
1	23	5	31	4	35	3	23	5	27	4	35	3
2	23	6	27	4	35	3	23	6	27	4	35	3
3	23	6	31	5	35	3	23	6	27	5	35	3
4	27	6	31	5	35	3	23	6	27	5	35	4

Table 4 shows the choice probabilities associated with the designs. The ranges of the RUM and P-RRM choice probabilities are by and large the same for both efficient designs. From the choice tasks (Table 3), or the choice probabilities (Table 4) it is hard (if not impossible) to tell which design is efficient given either presupposed decision rule, and which is not so efficient.

Table 4: Choice probabilties of efficient designs

	RUM DESIGN			P-RRM DESIGN		
	P(Y=A)	**P(Y=B)**	**P(Y=C)**	**P(Y=A)**	**P(Y=B)**	**P(Y=C)**
RUM						
1	0.33	0.27	0.40	0.27	0.40	0.33

2	0.12	0.48	0.40	0.12	0.48	0.40
3	0.19	0.16	0.65	0.17	0.26	0.57
4	0.12	0.17	0.71	0.27	0.40	0.33
P-RRM						
1	0.24	0.41	0.36	0.22	0.56	0.22
2	0.07	0.67	0.26	0.07	0.67	0.26
3	0.18	0.30	0.52	0.18	0.46	0.35
4	0.13	0.32	0.55	0.22	0.56	0.22

To further explore the designs, Figure 6 visualizes the choice tasks for both designs. The black points represent alternatives. Alternatives within the same choice task are connected by blue lines. The outer left plot shows all 16 possible choice tasks. All blue lines have a negative slope. This is a result of the fact that we removed choice tasks containg dominant alternatives. The middle plot shows the 4 choice tasks belonging to the most efficient RUM design; the outer right plot shows the 4 choice tasks belonging to the most efficient RRM design.

Figure 6 reveals that the RUM and RRM designs are quite unsimilar, albeit they do have one choice task in common. Not either from the visualisations it is trivial to see why one design would be efficient for one decision rule, while relatively inefficient for the other. Noteworthily, in the RRM design the all four blue lines – connecting the alternatives within a choice task – have a convex shape. In contrast, in the RUM design, just one 3 line is convex, the other three being concave. Furthermore, it can be noticed that in the RUM design all 4 attribute levels for both travel cost and travel time are used, while the RRM design uses only 3 levels for the travel time attribute.

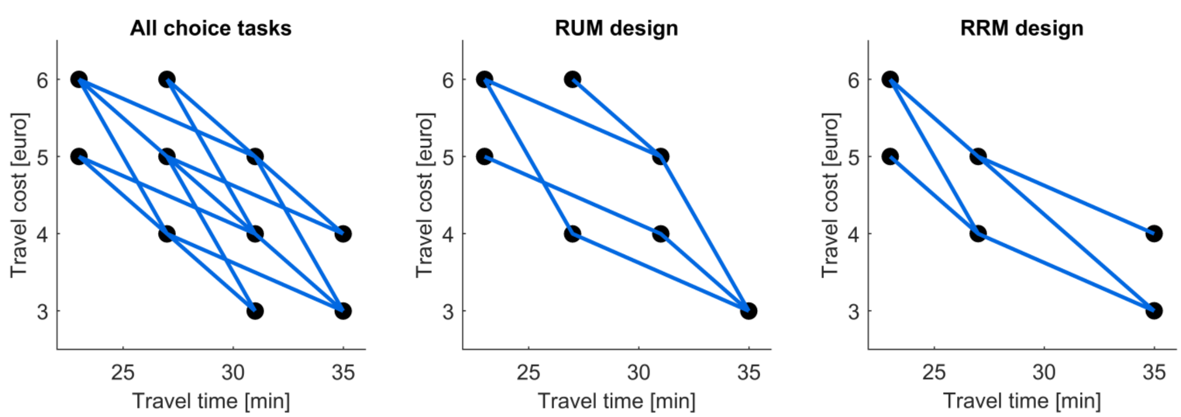


Figure 6: Visualisation of choice tasks (designs without dominant alternatives)

Out of methodological interest we further investigate the differences between RUM and RRM efficient designs using the all possible choice tasks, instaed of only those choice tasks that do not contain dominant alternatives. To do so, we computed the D-error statistics for all ≈ 4×10^9 designs and selected those designs having the lowest D-error statistic. Table 5 shows the choice tasks of the optimal RUM design (left) and P-RRM design (right); Table 6 shows the choice probabilities associated with the optimal designs; and Figure 7 visualizes the choice tasks.

Four things catch the eye. Firstly, the designs are considerably more efficient than those generate constructed using only choice tasks that do not contain dominant alternatives. The D-error statistic for the RUM design is about four times smaller; the D-error statistic for the RRM design is two times smaller. Secondly, in both designs all four choice tasks contain dominant alternatives. Therefore, from a practical point of view these designs are not useful, despite their substantial improvements in statistical efficiency. Thirdly, the designs are very unsimilar in terms of the choice tasks. Finally, from the choice tasks, or the associated probabilities, it is hard to uncover which design is more efficient, given a pre-supposed decision rule.

Table 5: Experimental designs

	RUM DESIGN (100-12227282)						P-RRM DESIGN (127-5992943)					
D_{RUM}	0.0047						0.0067					
D_{RRM}	0.0106						0.0054					
Choice	**Route A**		**Route B**		**Route C**		**Route A**		**Route B**		**Route C**	
task	**TT**	**TC**	**TT**	**TC**	**TT**	**TC**	**TT**	**TC**	**TT**	**TC**	**TT**	**TC**
1	23	3	35	3	35	4	23	4	23	6	35	3
2	23	5	35	5	35	6	23	5	23	6	35	3
3	23	6	27	6	35	3	23	5	27	6	35	3
4	23	6	35	3	35	6	23	6	31	3	35	5

Table 6: Choice probabilties of efficient designs

	RUM DESIGN			P-RRM DESIGN		
	P(Y=A)	**P(Y=B)**	**P(Y=C)**	**P(Y=A)**	**P(Y=B)**	**P(Y=C)**
RUM						
1	0.82	0.13	0.05	0.63	0.09	0.28
2	0.82	0.13	0.05	0.39	0.14	0.47
3	0.21	0.11	0.68	0.41	0.08	0.50

4	0.22	0.74	0.04	0.13	0.81	0.06
P-RRM						
1	0.72	0.22	0.06	0.80	0.06	0.14
2	0.72	0.22	0.06	0.62	0.16	0.21
3	0.37	0.25	0.37	0.59	0.10	0.30
4	0.28	0.63	0.09	0.12	0.79	0.09

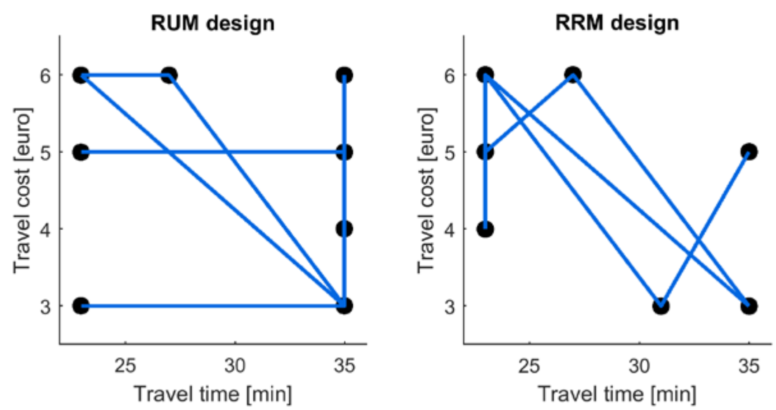


Figure 7: Visualisation of choice tasks

4.2 Data collection

The data collection took place in The Netherlands in May 2016. Respondents were recruited using a panel company (TNS NIPO). Only respondents who commuted two or more days per week by car were admitted. In total 106 respondents completed the full survey. Socio-demopgraphic characteristics of the respondents were provided by the panel company, although for 17 respondents socio-demographic data were missing due to registration errors on the side of the respondents. Table 7 shows the sample statistics. As can be seen, a relatively balanced sample has been obtained in terms of gender, age, eductation and income.

Each respondent is given both the RUM and the P-RRM SC experiment. The first choice task that is presented is the choice task 2 (see Table 3) as this choice task is belongs to both the RUM and RRM design. To avoid any ordering effects, the remaining six choice tasks belonging to the RUM and P-RRM designs were presented alternately in a random order. Also the order of the alternatives within each choice task was shuffled to reduce the sense of repetition on the side of the respondent. Finally, as in our contract with the panel company we were entitled to give respondents 10 choice tasks, three additional choice tasks were added (from the 9 choice tasks that were left that did not contain dominant alternatives).

Table 7: Sample statistics

Variable	Sample frequency	Percentage [%] in sample
Gender		
Male	44	42%
Female	45	42%
Missing	17	16%
Age		
18 to 24 yr.	2	2%
25 to 34 yr.	24	23%
35 to 44 yr.	25	24%
45 to 54 yr.	21	20%
55 to 64 yr.	15	14%
65 to 74 yr.	2	2%
Missing	17	16%
Completed education		
No education	0	0%
Elementary school	7	7%
Lower education	5	5%
Middle education	39	37%
Higher education	34	32%
University education	4	4%
Missing	17	16%
Income		
I < €9,400	2	2%
€9,400 ≤ I < €14,700	4	4%
€14,700 ≤ I < €20,600	5	5%
€20,600 ≤ I < €33,500	14	13%
€33,500 ≤ I < €67,000	37	35%
I ≥ €67,000	27	25%
Missing	17	16%

4.3 Results

Table 8 shows estimation results based on 1) the RUM Design data only, 2) the P-RRM Design dat only, and 3) the data of both designs. For each data set we estimated a RUM-MNL and a P-RRM-MNL model.

Based on Table 8 a number of observations can be made. Firstly, the results suggest that the model fit and the design decision rule are not independent. Specifically, we see that the RUM model obtains the best model fit when the design decision rule is RUM, while the P-RRM model obtains the best model fit when the design decision rule is P-RRM. In fact, the Ben-Akiva-Swait test (Ben-Akiva and Swait 1986) for nonnested models shows that the model fit differences are highly significant (p = 0.00). These result adds to the growing concern of bias cause by efficient SC experimental designs (e.g. Fosgerau and Börjesson 2015; Ojeda-Cabral et al. 2016b).

Secondly, in line with expectations, we see that the RUM model attains the lowest empirical D-error[8] (and hence smallest standard errors) when estimated based on the RUM design data, while the P-RRM-MNL model attains the lowest empirical D-error when estimated on the P-RRM design data. Specifically, in case the design decision rule matches the model decision rule the D-error is about twice as small than in case it does not.

Thirdly, the ratios of the parameters of the same model are significantly different from one another across the two data sets. For instance, the ratio of β_{time} over β_{cost} of the RUM model when estimated on RUM design data is 0.276 (implying a VoT of €16.56 per hour), while when estimated on the P-RRM design data it equals 0.221 (implying a VoT of €13.26 per hour). An independent sample t-test shows that these values are highly significantly different from one another (p = 0.00). This suggests that VoTs reported in numerous previous studies based on efficient designs might be biased, or at least influenced, by the efficient design itself.

Finally, as indicated by the goodness-of-fit measure ρ^2 the explanatory power of the models is generally rather low (i.e. $\rho^2 < 0.10$). However, given that we estimate straightforward MNL models that do not account for unobserved taste heterogeneity or panel effects, this is to be expected.

Table 8: Estimation results

[8] We use 'empirical D-error' to refer to the determinant of the estimated AVC.

DATA SET	RUM data				P-RRM data				All data			
Individuals	106				106				106			
Observations	424				424				742			
LL(0)	-465.8				-465.8				-815.2			
RUM-MNL												
LL(β)	-437.2				-448.9				-786.2			
ρ^2	0.06				0.04				0.04			
B_time	-0.196	0.026	-7.49	0.00	-0.187	0.033	-5.66	0.00	-0.163	0.022	-7.47	0.00
B_cost	-0.711	0.104	-6.82	0.00	-0.848	0.156	-5.45	0.00	-0.651	0.094	-6.92	0.00
Ratio	0.276	0.016	16.9	0.00	0.221	0.012	17.8	0.00	0.250	0.014	18.5	0.00
Empirical D-efficiency	0.0135				0.0306				0.0062			
P-RRM-MNL												
LL(β)	-448.7				-444.3				-795.0			
ρ^2	0.04				0.05				0.02			
B_time	-0.137	0.024	-5.74	0.00	-0.098	0.015	-6.35	0.00	-0.094	0.015	-6.25	0.00
B_cost	-0.460	0.093	-4.97	0.00	-0.497	0.085	-5.88	0.00	-0.385	0.071	-5.44	0.00
Ratio	0.298	0.031	9.72	0.00	0.197	0.022	8.91	0.00	0.244	0.023	10.37	0.00
Empirical D-efficiency	0.0134				0.0071				0.0035			

5 How does the design decision rule influences the modelling results?

The results presented in section 4.3 suggest that the design decision rule may influence the modelling results. This section further explores this conjecture. To do so, we explode our data set and analyse every possible design containing 4 choice tasks. In total $\binom{10}{4} = 210$ designs are constructed. First, we investigate the relation between RUM and P-RRM D-errors (section 5.1), then we investigate the differences in model fit between RUM and P-RRM based of these 210 designs (section 5.2), and finally we explore the relationship between the statistical efficiency and the model fit differences (section 5.3).

5.1 Relation between RUM and P-RRM efficiency

Figure 8 shows the relation between the RUM and P-RRM statistical efficiency. The left-hand scatter plot depicts the relation between the empirical D-error (thus, based on the determinants of the AVCs of the estimated models). The plot show that there is no particular relation between the RUM D-error and the P-RRM D-error. This result again confirms the view that efficient design are not robust towards misspecification of the underlying decision rule. The middle scatter plot depicts the relation between the D-error based on the initial priors (β_{Time} = -0.15, β_{Cost} = -1.00). Because, in retrospect, the priors used to construct the efficient design were chosen too large it is necessary to see whether these results are roughly the same as those in the left-hand plot. Finally, the right-hand side scatter plot shows the relation between the efficiencies of the designs based on the priors and the attained empirical efficiencies. It shows that although the priors were chosen too small, there is a strong relation between the efficiency based on the priors and the attained empirical efficiency. This indicates that our results are not driven by the fact that the priors do not coincide with the parameter estimates. Moreover, it implies that the analytical work in section 3 and these empirical analyses are 'consistent'.

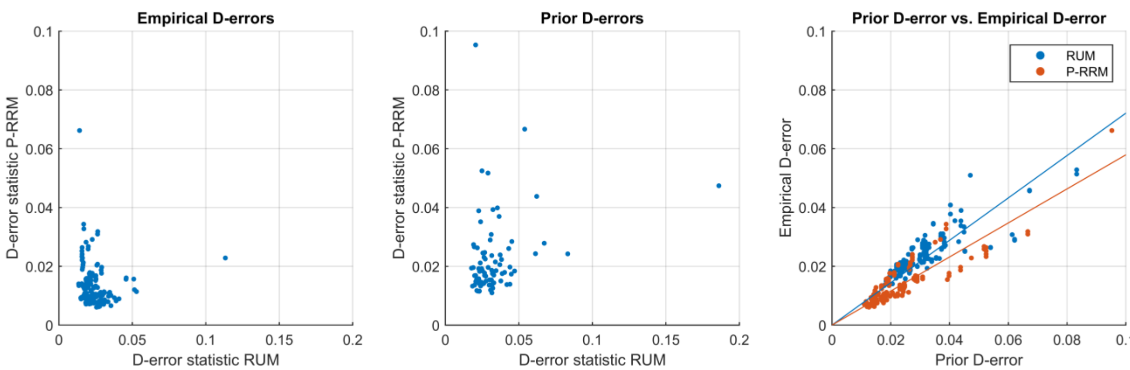


Figure 8: Relation between RUM and P-RRM efficiency

5.2 Relation between RUM and P-RRM Model fit

Figure 9 shows the relation between the attained model fits of the RUM and P-RRM model. The left-hand plot shows a scatter plot having the LL of the RUM on the x-axis and the LL of the P-RRM model on the y-axis. It shows that there is a strong correlation between the LL of the RUM model and the LL of the P-RRM model. This is likely the result of the fact that one set of choice tasks is easier than another – which typically leads to higher choice consistency under both models. Furthermore, about half of the points lie above the line $y = x$, implying that in half of the designs the P-RRM model yields the best model fit, while in the other half of the design the RUM models performs best. The right-hand plot depicts a histogram of the difference in LL between the RUM and the P-RRM model. It shows that the model fit difference has a bimodal distribution. While both LL_{RUM} and $LL_{P\text{-}RRM}$ are approximately normally distributed, due to the strong correlation between the two LLs, their difference is a bimodal distribution. Furthermore, the right-hand side plot shows that the model fit differences can be substantial. In over half of the designs the (absolute) model fit difference is larger than 5 LL points. This result signals that inferences on what decision best describes the decision process are highly dependent on the design that is being used for the analysis. In other words, based on one design a researcher may conclude that a RUM model explains the choice data significantly better than a P-RRM model, while the exact opposite conclusion could be reached when another design would have been used.

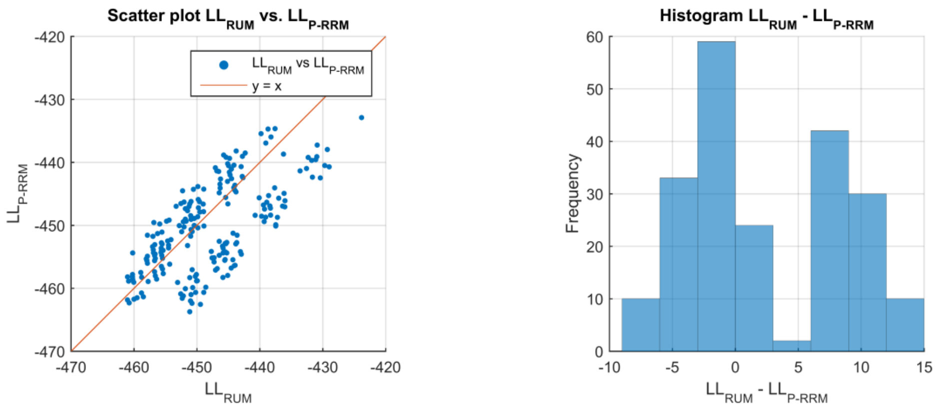


Figure 9: Relation between RUM and P-RRM model fit

5.3 Does the design decision rule biases the modelling outcomes?

To investigate whether the design decision rule systematically biases the modelling outcomes – in the sense that inferences made by the analyst on what is the most likely decision rule that is being employed by the decision makers, or what is there VoT are misguided. To do so, we explore whether there exist a correlation between the efficiency conditional on a decision rule (RUM or P-RRM) and the model fit difference between the RUM and the P-RRM model, and whether there exist a correlation between the efficiency conditional on a decision rule and the implied VoT.

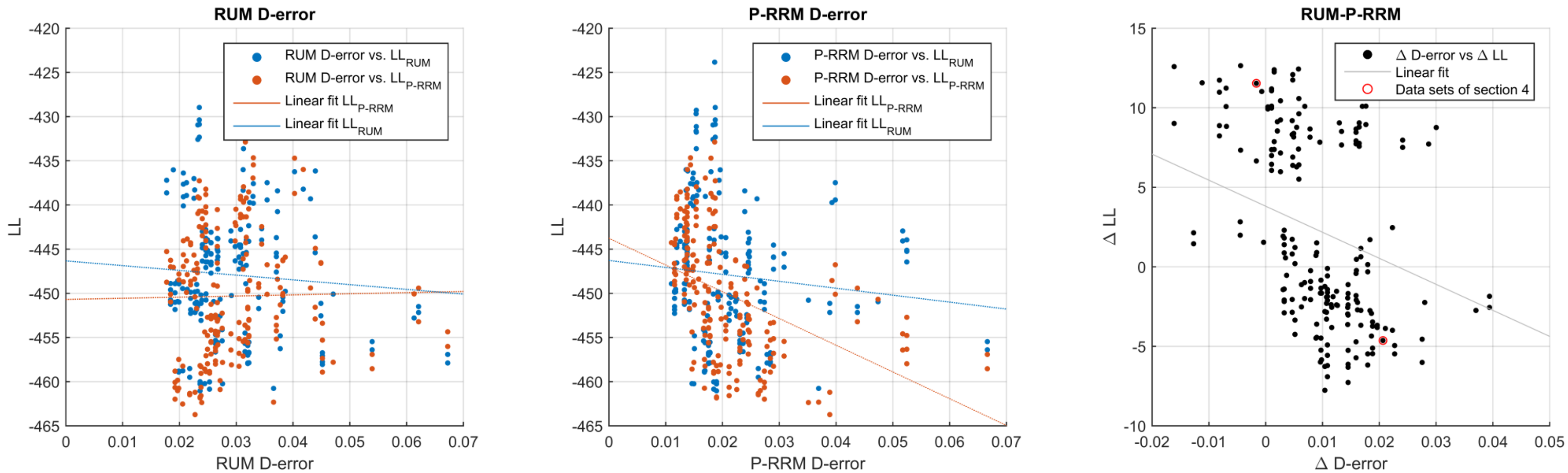
5.3.1 Does the design decision rule biases model fit?

Figure 10 displays scatter plots showing the relation between the D-error and model fit. The left-hand side plot depicts the relation between the RUM D-error and the RUM and P-RRM model fits. The middle plot depicts the relation between the P-RRM D-error and the RUM and P-RRM model fits. The right-hand side plot shows the relation between the difference in RUM and P-RRM D-error and the difference in RUM and P-RRM model fit. In all plots linear regression lines are added to see the general tendencies. Furthermore, below each plot Pearson correlation coefficients are given, showing the direction of the relation and whether the relation is statistically significant.

Based on Figure 10 a number of inferences can be made. Firstly, the negative slopes of the blue line in the left-hand plot, and the orange line in the middle plot reveal that in case the model and the design decision rule match, a smaller D-error (higher efficiency) is associated with an on average higher model fit. The pearson correlation coefficients show that this relation is statistically highly significant for the P-RRM D-error and the P-RRM model (ρ = -0.47, p = 0.00). However, for the RUM D-error and the RUM model this relation is not statistically significant at conventional levels of significance (ρ = -0.10, p = 0.14). Yet, in case the model and the design decision rule do not match, we find no statistically significant relation between D-error and model fit.

Secondly, the right-hand side plot shows that in case a design is efficient for RUM while relatively inefficient for P-RRM the researcher is likely to find the RUM model to outperform the P-RRM model, and vice versa. The results of section 4 are just a typically manifestation of this relation. To show this, the two data sets used for analysis in section 4 are highlighted in red in the right-hand side plot.

All in all, these results point in the direction that there exist a relation between statistical efficiency and model fit of the model that matches the design decision rule. This relation is identified to be highly significant for P-RRM design data , but not for RUM design data. However, as – to the best of our knowledge – there is no good reason to suspect that the relation between efficiency and model fit would asymmetrically hold (thus only for P-RRM design data but not for RUM design data) these results imply that researcher should be very cautious when drawing conclusion on what is the most likely decision rule that is being employed by the decision makers based on efficient SC data.



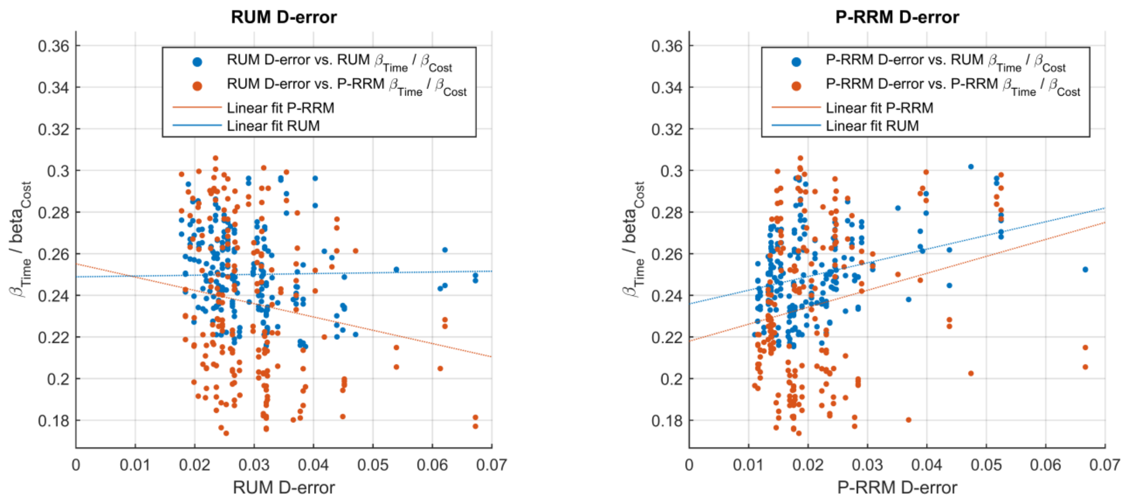
	ρ	p-value
RUM	-0.10	0.14
P-RRM	0.03	0.70

	ρ	p-value
RUM	-0.11	0.11
P-RRM	-0.47	0.00

ρ	p-value
-0.43	0.00

Figure 10: D-error vs model fit

5.3.2 *Does the design decision rule biases behavioural insights?*



	ρ	p-value
RUM	0.03	0.67
P-RRM	-0.24	0.00

	ρ	p-value
RUM	0.37	0.00
P-RRM	0.23	0.00

Figure 11: D-error vs parameter ratio

6 Conclusion and discussion

Appendix A: Second order derivative of the Log-Likelihood function w.r.t. its parameters for the P-RRM model

This appendix derives the second order derivative of the Log-Likelihood function of the P-RRM model w.r.t. it parameters. The second order derivatives are needed to construct the fisher information matrix, which in turn is needed to evaluate the statistical efficiency of a design.

$$\frac{\partial^2 LL^{P-RRM}}{\partial\beta_{m_1}\partial\beta_{m_2}} = -\frac{\partial}{\partial\beta_{m_2}}\left(\sum_{s=1}^{S}\sum_{j=1}^{J} y_{js} - P_{js}\right)\tilde{x}_{jms}^{P-RRM}$$

$$\frac{\partial^2 LL^{P-RRM}}{\partial\beta_{m_1}\partial\beta_{m_2}} = \sum_{s=1}^{S}\sum_{j=1}^{J}\tilde{x}_{jms}^{P-RRM}\ \frac{\partial}{\partial\beta_{m_2}}P_{js}$$

$$\frac{\partial^2 LL^{P-RRM}}{\partial\beta_{m_1}\partial\beta_{m_2}} = \sum_{s=1}^{S}\sum_{j=1}^{J}\tilde{x}_{jms}^{P-RRM}\ \frac{\partial}{\partial\beta_{k_2}}\left(\frac{e^{-\sum_{m=1}^{M}\beta_m x_{jms}^{P-RRM}}}{\sum_{i=1}^{J} e^{-\sum_{m=1}^{M}\beta_m x_{ims}^{P-RRM}}}\right)$$

Applying the quotient rule:

$$f = e^{-\sum_{m=1}^{M}\beta_m x_{jms}^{P-RRM}}, \qquad \frac{\partial f}{\partial\beta_{m_2}} = -\tilde{x}_{jm_2s}^{P-RRM}\cdot e^{-\sum_{m=1}^{M}\beta_m x_{jms}^{P-RRM}}$$

$$g = \sum_{i=1}^{J} e^{-\sum_{m=1}^{M}\beta_m x_{ims}^{P-RRM}}, \qquad \frac{\partial g}{\partial\beta_{m_2}} = -\sum_{i=1}^{J} x_{im_2s}^{P-RRM}\cdot e^{-\sum_{m=1}^{M}\beta_m x_{ims}^{P-RRM}}$$

$$\frac{\partial^2 LL^{P-RRM}}{\partial\beta_{m_1}\partial\beta_{m_2}} = \sum_{s=1}^{S}\sum_{j=1}^{J}\tilde{x}_{jms}^{P-RRM}\left(\frac{\left[-\tilde{x}_{jm_2s}^{P-RRM}\cdot e^{-\sum_{m=1}^{M}\beta_m x_{jms}^{P-RRM}}\right]\left[\sum_{i=1}^{J} e^{-\sum_{m=1}^{M}\beta_m x_{ims}^{P-RRM}}\right]-\left[e^{-\sum_{m=1}^{M}\beta_m x_{jms}^{P-RRM}}\right]\left[-\sum_{i=1}^{J} x_{im_2s}^{P-RRM}\cdot e^{-\sum_{m=1}^{M}\beta_m x_{ims}^{P-RRM}}\right]}{\left(\sum_{i=1}^{J} e^{-\sum_{m=1}^{M}\beta_m x_{ims}^{P-RRM}}\right)^2}\right)$$

Rearranging terms, yields equation X and X

$$\frac{\partial^2 LL^{P-RRM}}{\partial\beta_{m_1}\partial\beta_{m_2}} = \sum_{s=1}^{S}\sum_{j=1}^{J}\tilde{x}_{jms}^{P-RRM}\left(\frac{-\tilde{x}_{jm_2s}^{P-RRM}\cdot e^{-\sum_{m=1}^{M}\beta_m x_{jms}^{P-RRM}}}{\left(\sum_{i=1}^{J} e^{-\sum_{m=1}^{M}\beta_m x_{ims}^{P-RRM}}\right)} - \frac{e^{-\sum_{m=1}^{M}\beta_m x_{jms}^{P-RRM}}}{\sum_{i=1}^{J} e^{-\sum_{m=1}^{M}\beta_m x_{ims}^{P-RRM}}}\cdot\frac{-\sum_{i=1}^{J} x_{im_2s}^{P-RRM}\cdot e^{-\sum_{m=1}^{M}\beta_m x_{ims}^{P-RRM}}}{\sum_{i=1}^{J} e^{-\sum_{m=1}^{M}\beta_m x_{ims}^{P-RRM}}}\right)$$

$$\frac{\partial^2 LL}{\partial\beta_{m_1}\partial\beta_{m_2}} = -\sum_{s=1}^{S}\sum_{j=1}^{J}\tilde{x}_{jms}^{P-RRM}\left(\tilde{x}_{jm_2s}^{P-RRM} P_{js} - P_{js}\cdot\sum_{i=1}^{J} x_{im_2s}^{P-RRM} P_{is}\right)$$

Finally, we obtain the second order derivative of the Log-Likelihood function w.r.t. the generic model parameters β_1 and β_2 (equation X)

$$\frac{\partial^2 LL}{\partial\beta_{m_1}\partial\beta_{m_2}} = -\sum_{s=1}^{S}\sum_{j=1}^{J}\tilde{x}_{jms}^{P-RRM} P_{js}\left(\tilde{x}_{jm_2s}^{P-RRM} - \sum_{i=1}^{J} x_{im_2s}^{P-RRM} P_{is}\right)$$

References

Abrantes, P. & Wardman, M. (2009). Meta-analysis of UK values of time: an update. European Transport Conference, 2009.

Ben-Akiva, M. & Swait, J. (1986). The Akaike Likelihood Ratio Index. *Transportation Science, 20(2)*, 133-136.

Bliemer, M. C. & Rose, J. M. (2010). Construction of experimental designs for mixed logit models allowing for correlation across choice observations. *Transportation Research Part B: Methodological*, *44(6)*, 720-734.

Bliemer, M. C., Rose, J. M. & Hensher, D. A. (2009). Efficient stated choice experiments for estimating nested logit models. *Transportation Research Part B: Methodological*, *43(1)*, 19-35.

Bliemer, M. C. J. & Rose, J. M. (2011). Experimental design influences on stated choice outputs: An empirical study in air travel choice. *Transportation Research Part A: Policy and Practice*, *45(1)*, 63-79.

Bliemer, M. C. J., Rose, J. M. & Chorus, C. G. (2016). Detecting dominancy in stated choice data and accounting for dominancy-based scale differences in logit models. *Transportation Research Part B*.

Boeri, M., Scarpa, R. & Chorus, C. G. (2014). Stated choices and benefit estimates in the context of traffic calming schemes: Utility maximization, regret minimization, or both? *Transportation Research Part A: Policy and Practice*, *61*, 121-135.

Börjesson, M. & Eliasson, J. (2014). Experiences from the Swedish Value of Time study. *Transportation Research Part A: Policy and Practice*, *59*, 144-158.

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.

Chorus, C. (2012a). Random Regret Minimization: An Overview of Model Properties and Empirical Evidence. *Transport Reviews*, *32(1)*, 75-92.

Chorus, C. G. (2010). A new model of random regret minimization. *European Journal of Transport and Infrastructure Research*, *10(2)*, 181-196.

Chorus, C. G. (2012b). Random regret-based discrete choice modeling a tutorial. SpringerBriefs in business,. Berlin ; New York, Springer,**:** 1 online resource (viii, 55 p.).

Chorus, C. G., Arentze, T. A. & Timmermans, H. J. P. (2008). A random regret-minimization model of travel choice. *Transportation Research Part B-Methodological*, *42(1)*, 1-18.

Daly, A., Tsang, F. & Rohr, C. (2014). The value of small time savings for non-business travel. *Journal of Transport Economics and Policy (JTEP)*, *48(2)*, 205-218.

Dekker, T. (2014). Indifference based value of time measures for Random Regret Minimisation models. *Journal of Choice Modelling*, *12(0)*, 10-20.

Ferrini, S. & Scarpa, R. (2007). Designs with a priori information for nonmarket valuation with choice experiments: A Monte Carlo study. *Journal of Environmental Economics and Management*, *53(3)*, 342-363.

Fosgerau, M. & Börjesson, M. (2015). Manipulating a stated choice experiment. *Journal of Choice Modelling*, *16*, 43-49.

Hensher, D. A., Barnard, P. O. & Truong, T. P. (1988). The role of stated preference methods in studies of travel choice. *Journal of Transport Economics and Policy*, 45-58.

Hess, S., Bierlaire, M. & Polak, J. W. (2005). Estimation of value of travel-time savings using mixed logit models. *Transportation Research Part A: Policy and Practice*, *39(2–3)*, 221-236.

Hess, S., Stathopoulos, A. & Daly, A. (2012). Allowing for heterogeneous decision rules in discrete choice models: an approach and four case studies. *Transportation*, *39(3)*, 565-591.

Huber, J., Payne, J. W. & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of consumer research*, *9(1)*, 90-98.

Huber, J. & Zwerina, K. (1996). The Importance of Utility Balance in Efficient Choice Designs. *Journal of Marketing Research*, *33(3)*, 307-317.

Kessels, R., Jones, B., Goos, P. & Vandebroek, M. (2011). The usefulness of Bayesian optimal designs for discrete choice experiments. *Applied Stochastic Models in Business and Industry*, *27(3)*, 173-188.

Kouwenhoven, M., de Jong, G. C., Koster, P., van den Berg, V. A. C., Verhoef, E. T., Bates, J. & Warffemius, P. M. J. (2014). New values of time and reliability in passenger transport in The Netherlands. *Research in Transportation Economics*, *47*, 37-49.

Louviere, J. J., Hensher, D. A. & Swait, J. D. (2000). *Stated choice methods : analysis and applications*. (Cambridge, UK: Cambridge University Press).

Ojeda-Cabral, M., Batley, R. & Hess, S. (2016a). The value of travel time: random utility versus random valuation. *Transportmetrica A: Transport Science*, *12(3)*, 230-248.

Ojeda-Cabral, M., Hess, S. & Batley, R. (2016b). Understanding valuation of travel time changes: are preferences different under different stated choice design settings? *Transportation*, 1-21.

Palhazi Cuervo, D., Goos, P. & Sörensen, K. (2014). Optimal design of large-scale screening experiments: a critical look at the coordinate-exchange algorithm. *Statistics and Computing*, 1-14.

Rose, J. M. & Bliemer, M. C. J. (2009). Constructing Efficient Stated Choice Experimental Designs. *Transport Reviews: A Transnational Transdisciplinary Journal*, *29(5)*, 587-617.

Rose, J. M. & Bliemer, M. C. J. (2013a). Incorporating analyst uncertainty in model specification of respondent processing strategies into efficient designs for logit models. The 59th World Statistics Congress, 25-30 August 2013, hong Kong.

Rose, J. M. & Bliemer, M. C. J. (2013b). Sample size requirements for stated choice experiments. *Transportation*, *40(5)*, 1021-1041.

Rose, J. M. & Bliemer, M. C. J. (2014). Stated choice experimental design theory: the who, the what and the why. In (Eds.), *Handbook of Choice Modelling* (pp. Cheltenham, UK: 'Edward Elgar Publishing, Inc.'.

Sandor, Z. & Wedel, M. (2005). Heterogeneous Conjoint Choice Designs. *Journal of Marketing Research*, *42(2)*, 210-218.

Van Cranenburgh, S., Guevara, C. A. & Chorus, C. G. (2015). New insights on random regret minimization models. *Transportation Research Part A: Policy and Practice*, *74(0)*, 91-109.

Van Cranenburgh, S. & Prato, C. G. (2016). On the robustness of Random Regret Minimization modelling outcomes towards omitted attributes. *Journal of Choice Modelling,(18)*, 51-70.

Van Cranenburgh, S., Prato, C. G. & Chorus, C. G. (working paper) Accounting for variation in choice set size in Random Regret Minimization models,

Welch, M. & Williams, H. (1997). The Sensitivity of Transport Investment Benefits to the Evaluation of Small Travel-Time Savings. *Journal of Transport Economics and Policy*, *31(3)*, 231-254.

Yu, J., Goos, P. & Vandebroek, M. (2009). Efficient conjoint choice designs in the presence of respondent heterogeneity. *Marketing Science*, *28(1)*, 122-135.