Issue 21(3), 2021, pp.38-63
https://doi.org/10.18757/ejtir.2021.21.3.3949

TU Delft OPEN

EJTIR
ISSN: 1567-7141
http://ejtir.tudelft.nl/

What is really uncovered by mixing different model structures: contrasts between latent class and model averaging.

Thomas O. Hancock[1]
Choice Modelling Centre and Institute for Transport Studies, University of Leeds, UK.

Stephane Hess[2]
Choice Modelling Centre and Institute for Transport Studies, University of Leeds, UK.

Latent class models have long been a tool for capturing heterogeneity across decision-makers in the sensitivities to individual attributes. More recently, there has been increased interest in using these models to capture heterogeneity in actual behavioural processes, such as information/attribute processing and decision rules. This often leads to substantial improvement in model fit and the apparent finding of large clusters of individuals making choices in ways that are substantially different from those used by others. Such findings have however not been without criticism given the potential risk of confounding with other more model-specific heterogeneity. In this paper, we consider an alternative approach for exploring the issue by contrasting the findings obtained with model averaging, which combines the results from a number of separately (rather than simultaneously) estimated models. We demonstrate that model averaging can accurately recover the different data generation processes used to create a number of simulated datasets and thus be used to infer likely sources of heterogeneity. We then use this new diagnostic tool on two stated choice case studies. For the first, we find that the use of model averaging leads to significant reductions in the amount of heterogeneity of the type analysts have sought to uncover with latent class structures of late. For the second, results from model averaging show clear evidence of the existence of both taste and decision rule heterogeneity. Overall, however, our results suggest that heterogeneity in the sensitivities to individual attributes rather than the behavioural process per se could be the key factor behind the improvements gained through the adoption of latent class models for heterogeneity in behavioural processes.

Publishing history

Submitted: 3 October 2019
Accepted: 24 July 2021
Published: 31 August 2021

Cite as

Hancock, T.O., & Hess, S. (2021). What is really uncovered by mixing different model structures: contrasts between latent class and model averaging. *European Journal of Transport and Infrastructure Research, Issue 21(3)*, 38-63.

Keywords: *latent class, information processing, attribute non-attendance, decision rule heterogeneity*

[1] A: Choice Modelling Centre and Institute for Transport Studies, 34-40 University Road, University of Leeds, Leeds,LS2 9JT, UK.
E: T.O.Hancock@leeds.ac.uk
[2] A: Choice Modelling Centre and Institute for Transport Studies, 34-40 University Road, University of Leeds, Leeds,LS2 9JT, UK.
E: S.Hess@leeds.ac.uk

1. Introduction

Over the last decade, there has been increasing interest by choice modellers to allow for departures from traditional decision rules (Chorus, 2014) and/or the way in which individuals process the information describing the alternatives (Hensher, 2014). Much of the work has looked at contrasts between models using one specific alternative decision rule or process to the results against an alternative model, i.e. fitting the same process to an entire sample of decision-makers. However, a growing number of studies (Hess and Rose, 2007; Scarpa et al., 2009; Hensher and Greene, 2010; Campbell et al., 2010; Hole, 2011; Hensher et al., 2012; Hess et al., 2012; Charoniti et al., 2020; Rezapour and Ksaibati, 2021; Smith et al., 2021) have also looked at allowing for heterogeneity in the actual underlying model structure across individuals in a single sample. This work has mainly made use of latent class (LC) structures, with two key applications, namely decision rule heterogeneity (where the different classes within a latent class model adopt different decision rules) and in information processing work (where typically the same model is used in the different classes but with different attributes included). In both of these applications, the key idea is that each of the classes will better capture the choices of some share of the decision-makers.

While the work using latent class structures for heterogeneity in either decision rules or information processing strategies has been shown to lead to substantial improvement (Hess and Rose, 2007; Hess et al., 2012) in fit and apparent meaningful insights, it has also not been without criticism. In particular, concerns have been raised about the extensive risk of confounding between heterogeneity in the sensitivities to individual attributes and heterogeneity in the process or model structure.

In a traditional latent class model, the different β parameters in different classes are used solely to uncover taste heterogeneity. In a latent class model that combines different structures in different classes, these individual models will themselves be making use of different β parameters, while in the case of attribute non-attendance (ANA), they will use different combinations of the β parameters. For reasons of complexity, the vast majority of applications have used just a single class per behavioural process, whether that be one class for each decision rule (e.g. Random Utility Models (RUM), Random Regret Minimisation (RRM), etc) or one class for each of the combinations of considered attributes in an ANA context. Maximum likelihood estimation will simply converge to those parameters that give the best mathematical fit to the data. For example, imagine a situation where the decisions of all individuals in the data are best explained by a RUM structure, but where there are variations across individuals in the sensitivity to for example the cost attribute. If the analyst estimates a LC model with two classes, where one class uses a RUM model and the other class uses a RRM model, then the only mechanism available to maximum likelihood process for explaining the heterogeneity in cost sensitivities is to allocate a non-zero class allocation probability to both the RUM and RRM classes, with different cost coefficients in the two models. In other words, even in the absence of decision rule heterogeneity, the model will *uncover* such heterogeneity if the benefit of being able to use different cost sensitivities in the RUM and RRM classes outweighs the loss in fit of using a RRM model to explain the choices of people who made decisions more in line with RUM. There is thus the real possibility that apparent evidence of decision rule heterogeneity will be driven by heterogeneity in sensitivities rather than actual decision process.

These concerns have found empirical support in the work of Hess et al. (2013b) who show that the share for non-attendance classes reduces substantially when allowing for additional random heterogeneity, while the work of Hess et al. (2016) shows that allowing for random heterogeneity in the parameters of RUM and RRM models within a RUM-RRM mixture model substantially reduces the extent of decision rule heterogeneity. The use in practice of such latent class models allowing for different structures in different classes continues to be very popular (Boeri and Longo, 2017; Dey et al., 2018) despite these concerns. A key reason is likely that the inclusion of additional taste heterogeneity (moving from finite latent class models to continuous mixture models), as in the work of Hess et al. (2013b) and Hess et al. (2016) is computationally very difficult. The same applies to the inclusion of additional classes with models of the same type (e.g. using two RUM classes and two RRM classes), where this also leads to a proliferation in the number of parameters. Whilst there are of course other methods that can be adopted to explore these and other kinds of heterogeneity, the key aim of the present paper is to specifically consider a different approach to further examine the results of these latent class models which are so popular, but without increasing computational demands. We do this by highlighting how model averaging can be used as a diagnostic tool for the potential confounding between taste heterogeneity and other heterogeneity highlighted in these models.

Model averaging uses a sequential latent class approach, estimating first the individual candidate models at the sample level, before then combining these models in a latent class model which keeps the model-specific parameters fixed and only estimates the model weights. We illustrate this process on simulated data as well as typical stated preference (SP) data and show that model averaging can provide additional insights that could allow an analyst to reach a more informed decision as to the key drivers of heterogeneity in a model.

The aim of using model averaging in the present paper is to investigate potential cases of confounding in models using simultaneous estimation of different model structures. Of course, a caveat applies in that it is also possible that the presence of decision rule heterogeneity and/or heterogeneity in processing strategies can only be uncovered when estimating models in which the parameter estimates for the different subclasses are informed more by some individuals in the data than by others, as would be the case in simultaneous estimation. We address this point specifically by showing the possibility of including some models within the set M that themselves allow for heterogeneity. For example, it is straightforward to allow a given model m to be itself a LC or mixed multinomial logit (MMNL) structure. As the model will be estimated separately rather than as part of the overall model averaging structure, we avoid the computation issues of including complex structures within an overall latent class structure.

The remainder of this paper is organised as follows. We first summarise the methodology in Section 2. This is followed by a simulated data experiment demonstrating how model averaging can help recover the original data generation process (Section 3). The core empirical work follows in Section 4, where we look both at attribute non-attendance and decision rule heterogeneity for two stated preference (SP) case studies. Finally, some conclusions are presented in Section 5.

2. Methodology

Latent class (LC) structures have long been used as a tool for introducing heterogeneity across individual decision-makers in choice models (see Greene and Hensher, 2003; Hess, 2014, for background). In a latent class model, the population is probabilistically divided into S different classes, where the log-likelihood (LL) for the choices observed for a set of N decision-makers is given by:

$$LL(\beta,\pi) = \sum_{n=1}^{N} ln\left(\sum_{s=1}^{S}\pi_{ns}\prod_{t=1}^{T_n}P_{j^*_{nt}}(\beta_s)\right), \tag{1}$$

where j^*_{nt} is the actual alternative observed to be chosen by person n in situation t. We have that $\pi = \langle\pi_1,\ldots,\pi_N\rangle$, with π_n a vector whose element π_{ns} gives the share (probability) of individual n belonging to class s such that $\sum_{s=1}^{S}\pi_{ns} = 1$, $\forall n$ and $0 \geq \pi_{ns} \geq 1$, $\forall n,s$. These class allocation probabilities can vary across individual decision-makers as a function of their characteristics, using a class allocation model, such that $\pi_n = f(\gamma, z_n)$ where z_n are characteristics of person n and γ is a vector of estimated parameters.

In almost all applications of latent class models, $P_{j^*_{nt}}(\beta_s)$ is of the Multinomial Logit (MNL) type. Even when this was not the case, for example using Nested Logit (NL) models inside a LC structure, the focus for the first two decades of widespread use of LC models was very much on a case where the functional form of $P_{j^*_{nt}}(\beta_s)$ is the same across classes (i.e. $s = 1,\ldots,S$), with differences only in the parameters used in the classes, i.e. β_s in class s, where $\beta = \langle\beta_1,\ldots,\beta_S\rangle$. This use of latent class models thus focusses on capturing what would typically be called “taste heterogeneity" while maintaining homogeneity in the underlying behavioural process across individual decision-makers.

Latent class models have more recently been used for heterogeneity in decision rules and information processing. While the former has received more attention, the latter work actually takes historical precedence.

A key interest in the field of information processing strategies (IPS) or attribute processing strategies (APS) has been the notion that some decision-makers may actually make their choices based on only a subset of the attributes that describe the alternatives at hand. This phenomenon is typically referred to as attribute non-attendance (ANA) or attribute ignoring, and an in-depth review of work in this area is given in Hensher (2010). The interest in this topic in the present discussions comes in the context of ways to accommodate ANA in models. A key role in this area was played by the early discussions in Hess and Rose (2007), who proposed the use of a latent class approach to accommodate ANA, a method since adopted by numerous other studies (e.g. Scarpa et al., 2009; Hensher and Greene, 2010; Campbell et al., 2010; Hole, 2011; Hensher et al., 2012; Smith et al., 2021). With this approach, different latent classes relate to different combinations of attendance and non-attendance across attributes. For each attribute treated in this manner, there exists a non-zero coefficient (to be estimated), which is used in the *attendance classes*, while the attribute is not employed in the *non-attendance* classes, i.e. the coefficient is set to zero. In a complete specification, covering all possible combinations, this would thus lead to 2^K classes, with K being the number of attributes, where a given coefficient will take the same value in all classes where that attribute is included.

In addition to the vector β, we now have a SxK matrix Λ, in which each row, s, contains a different combination of 0 and 1 elements, where $S = 2^K$. Next, let $A \circ B$ be the element-by-element product of two equally sized vectors A and B, yielding a vector C of the same size, where the k^{th} element of C is obtained by multiplying the k^{th} element of A with the k^{th} element of B. Using this notation, the specific values used for the taste coefficients in class s are then given by the vector $\beta_s = \beta \circ \Lambda_s$. The k^{th} element of the vector β_s is thus the k^{th} element of β if $\Lambda_{s,k} = 1$, and zero otherwise. The log-likelihood is then given by:

$$LL(\beta,\pi) = \sum_{n=1}^{N} ln\left(\sum_{s=1}^{S} \pi_{ns} \prod_{t=1}^{T_n} P_{j^*_{nt}}\left(\beta_s = \beta \circ \Lambda_s\right)\right). \tag{2}$$

A different application of such heterogeneous structures in different classes has arisen in the context of decision rule heterogeneity. There has long been interest in the notion that different individuals make their decisions in different ways, going back to work in psychology in the 1970s (Montgomery and Svenson, 1976). Although structures belonging to the family of random utility models have come to dominate, it is important to recognise that alternative paradigms for decision-making have been proposed, for example the elimination by aspects model of Tversky (1972), but also more recent work based on the concepts of happiness (Abou-Zeid and Ben-Akiva, 2010) and regret (Chorus et al., 2008; Chorus, 2010). The evidence in the literature is that which paradigm works best is very much dataset specific.

Hess et al. (2012) put forward the hypothesis that variations in decision rules may be across decision-makers with a single dataset, not just across datasets, and propose the use of a confirmatory latent class (CLC) approach in this context. Specifically, let $P^{(m)}_{j^*_{nt}}(\beta_m)$ give the probability using a model of type m, with a vector of parameters β_m. The Hess et al. (2012) framework is based on the idea that different behavioural processes are used in the data. The original exposition by Hess et al. (2012) assumes that a different model type m is used in each class S, but this is not a requirement, and the same model structure could be used in more than one class. We then have:

$$LL(\beta,\pi) = \sum_{n=1}^{N} ln\left(\sum_{s=1}^{S} \pi_{ns} \prod_{t=1}^{T_n} P^{m_s}_{j^*_{nt}}\left(\beta_s\right)\right), \tag{3}$$

where m_s identifies the behavioural process used in class s, with β_s giving the vector of parameters used.

Hess et al. (2012) use the model to allow for mixtures between random utility maximisation, random regret minimisation (RRM) and elimination by aspects. They also discuss allowing for additional continuous random heterogeneity in parameters within individual classes, such that:

$$LL(\Omega,\pi)=\sum_{n=1}^{N} ln\left(\sum_{s=1}^{S}\pi_{ns}\int_{\beta_s}\prod_{t=1}^{T_n}P_{j^*_{nt}}^{m_s}(\beta_s)\,f(\beta_s\mid\Omega_s)\,\mathrm{d}\beta_s\right), \quad (4)$$

where $\beta_{ns}\sim f(\beta_{ns}\mid\Omega_s)$ and $\Omega=\langle\Omega_1,\ldots,\Omega_S\rangle$. In later work, Hess and Stathopoulos (2012) use an approach as in Walker and Ben-Akiva (2002) and Hess et al. (2013a), making the class allocation a function of a latent factor, which in this case also explains decision-makers' real world choices[1].

Model averaging, in this context, can be implemented as a sequential latent class model. Whereas a simultaneous model estimates the parameters of the class-specific models at the same time as the class allocation probabilities, a model averaging approach uses a sequential process. We first separately estimate the individual model from each class on the entire sample, before estimating the class allocation probabilities separately with the individual model parameters fixed. To apply model averaging, we thus first estimate a number of different individual models, where say $L_n^{(m)}(\Omega_m)$ gives the likelihood of the sequence of choices observed for person n, conditional on using model m, where this model uses a vector of parameters Ω_m. An analyst will estimate M different such models. Each model is estimated separately on the same data. Crucially, this implies that there needs to be some difference in the functional form between the different models, e.g. using different utility specifications, different mixing distributions, different attribute processing rules or indeed different decision rules. Indeed, any two models using the exact same structure will clearly converge to the same solution. Within-structure heterogeneity in sensitivities can easily be accommodated by some of the models being themselves LC or MMNL structures. In the context of the present paper, the set of M models would use different specifications for IPS or different specifications in terms of the underlying decision rules. The model averaging process then computes the overall likelihood for person n as the weighted average across M models, with the full sample log-likelihood given by:

$$LL(\Omega,\pi)=\sum_{n=1}^{N} ln\left(\sum_{m=1}^{M}\pi_{nm}L_n^{(m)}(\Omega_m)\right), \quad (5)$$

where $\sum_{m=1}^{M}\pi_{nm}=1$, $\forall n$ and $0\leq\pi_{nm}\leq=1$, $\forall n,m$. This overall log-likelihood is conditional on the vector of weights $\pi_n=\langle\pi_{n1},\ldots,\pi_{nM}\rangle$ for each person and the combined parameter estimates from the different models $\Omega=\langle\Omega_1,\ldots,\Omega_M\rangle$. Crucially, a sequential estimation process is used. The parameters Ω_m are estimated separately by maximising the log-likelihood only for model m, while the model weights are then estimated by maximising Equation 5 while keeping Ω fixed.

[1] At this stage, it should be noted that a latent class model mixing various decision rules is just one example of a wider set of structures that combine different models. A further possibility for example would be a model using different generalised extreme value (GEV) nesting structures in different latent classes, somewhat similar in aims to the work of Ishaq et al. (2013). Finally, a separate body of work looks at using different choice sets in different classes, in the context of choice set generation work (see e.g. Swait and Ben-Akiva 1985; Ben-Akiva and Boccara 1995 and Gopinath 1995, section 2.7).

3. Simulated data analysis

Before testing model averaging on our stated preference datasets, we use simulated datasets to look at the contrasting insights provided by different approaches when the true choice process is known. We first describe how we created our 11 different datasets and what types of heterogeneity each dataset contains. We then apply four different models corresponding to the four different decision rules used to create the different datasets, 10 different latent class models (4 with the same decision rule in both classes, and 6 with the different combinations of the four decision rules) and finally model averaging.

3.1 Generation of simulated data

We use four different decision rules to generate the choices in our simulated datasets:

1. A random utility model (RUM), where the utility for hypothetical respondent n in choice task t for alternative i is given by:

$$U_{nti} = \delta_i + \sum_{m=1}^{M} (\beta_m \cdot X_{ntim}) + \varepsilon_{nti}, \tag{6}$$

where β_m is the marginal utility coefficient associated with attribute m.

2. A regret minimisation model (RRM) based on the specification given by Chorus (2010), thus the regret is calculated:

$$R_{nti} = -\delta_i + \sum_{m=1}^{M} \sum_{j \neq i} ln(1 + exp(\beta_m \cdot (X_{ntjm} - X_{ntim}))) + \varepsilon_{nti}, \tag{7}$$

where the constants which are added to the regret are multiplied by -1 such that they have similar impacts on choice probabilities (as they do in RUM) by being applied in the same direction.

3. A pure regret minimisation model (P-RRM), as defined by van Cranenburgh et al. (2015), such that regret is calculated:

$$R_{nti} = -\delta_i + \sum_{m=1}^{M} \sum_{j \neq i} max(0, \beta_m \cdot (X_{ntjm} - X_{ntim})) + \varepsilon_{nti}. \tag{8}$$

4. A relative advantage maximisation model (RAM), as defined by Leong and Hensher (2014), which is equivalent to a random utility model with the addition of the comparison of relative advantages (RA) of alternative i with each of the other alternatives, meaning that the utility of the alternatives are now choice-set dependent:

$$U_{nti} = \delta_i + \sum_{m=1}^{M} (\beta_m \cdot X_{ntim}) + \sum_{j \neq i} RA_{ntij} + \varepsilon_{nti}, \tag{9}$$

where the relative advantage of an alternative i is calculated by comparing it against all other alternatives. The sum of advantages A_{ntij} is:

$$A_{ntij} = \sum_{m=1}^{M} ln(1 + exp(\beta_m \cdot (X_{ntjm} - X_{ntim}))). \tag{10}$$

With disadvantages $D_{ntij} = A_{ntji}$, the relative advantage is then estimated:

$$RA_{ntij} = \frac{A_{ntij}}{A_{ntij} + D_{ntij}}. \tag{11}$$

We use an efficient design to generate $5,000$ mode choice scenarios, each with four possible alternatives: car, air, rail and high-speed rail. These alternatives are described by (for respondent n, alternative i in choice scenario t) travel cost (TC_{nti}), travel time (TT_{nti}) and access time (AT_{nti}).

Thus, for the RUM model, the utility is specifically calculated:

$$U_{nti} = \delta_i + \delta_{F_i} \cdot z_{F,n} + \beta_{TT} \cdot \alpha_{TT_i} \cdot TT_{nti} + \beta_{TC} \cdot \alpha_{IE,n} \cdot TC_{nti} + \beta_{AT} \cdot AT_{nti} + \varepsilon_{nti}, \tag{12}$$

where δ_i and δ_{F_i} are alternative specific constants (with the constant for car normalised to zero), with δ_{F_i} only applying when the dummy variable for female respondents, $z_{F,n}$, equals one (which is the case for half of the participants). We have three marginal utility coefficients, β_{TT}, β_{TC}, β_{AT}, for travel time, travel cost and access time, respectively. We use car as the base for travel time sensitivity, and apply mode-specific multipliers for travel time sensitivity through α_{TT_i}. Finally, we incorporate an income effect, which is defined as $\alpha_{IE,n} = (\frac{I_n}{2500})^{\alpha_I}$, where I_n is the income for individual n and α_I is an income elasticity. Analogous specifications are used to calculate the regret for alternatives under RRM and P-RRM using Equations 7 and 8 respectively, and the utility under RAM with Equations 9-11.

For all decision rules, the assumption of type I extreme value errors for ε results in probabilities of choosing each alternative being generated using the well known Multinomial Logit (MNL) formula:

$$P_{nti} = \frac{exp(U_{nti})}{\sum_{j=1}^{4} exp(U_{ntj})}, \tag{13}$$

with U_{nti} being replaced by $-R_{nti}$ for RRM and P-RRM. For each dataset, we then use a set of uniform draws to select the ‘chosen’ alternatives in the 5,000 scenarios.

We use different sets of parameter values and decision rules to create datasets with and without taste heterogeneity and decision rule heterogeneity. We use two sets of true parameter values for RUM and P-RRM, and one set of true parameter values for RRM and RAM, with the parameter values given in Table 1. The first two datasets we create use the same data generation process (DGP) for all individuals and thus assume that all decision-makers use the same decision rule with the same coefficients to make their choices. The next two datasets are created such that there is taste heterogeneity, by using a random allocation such that half of the decision-makers use one DGP, and the other half use a different DGP, where both DGPs have the same decision rule but different parameter values. The next three datasets instead only have decision rule heterogeneity, by using DGPs with similar parameter values but different decision rules for the different individuals. The final four datasets contain both decision rule and taste heterogeneity. The DGPs used and types of heterogeneity included in each dataset are summarised in Table 2.

Table 1 : Coefficient values for the data generation processes, where the alternatives are car (C), rail (R), air (A) and high-speed rail (H), together with the choice shares for each mode. Note that $\delta_H = 0$, $\delta_{F_H} = 0$ and $\alpha_{TT_C} = 1$. Values are chosen for RUM1 and RUM2 such that they are very different. P-RRM1, P-RRM2, RRM1 and RAM1 have adjusted β-coefficients in comparison to RUM1 and RUM2 such that they have approximately the same scale. The other coefficients are unchanged.

Parameter	RUM1	P-RRM1	RRM1	RAM1	RUM2	P-RRM2
δ_C	-0.5000	-0.5000	-0.5000	-0.5000	1.0000	1.0000
δ_R	-1.5000	-1.5000	-1.5000	-1.5000	-0.5000	-0.5000
δ_A	-1.0000	-1.0000	-1.0000	-1.0000	1.0000	1.0000
δ_{F_C}	-0.5000	-0.5000	-0.5000	-0.5000	-0.2000	-0.2000
δ_{F_R}	0.5000	0.5000	0.5000	0.5000	1.5000	1.5000
δ_{F_A}	1.0000	1.0000	1.0000	1.0000	-0.5000	-0.5000
α_I	-0.5000	-0.5000	-0.5000	-0.5000	-0.3000	-0.3000
β_{TT}	-0.0040	-0.0020	-0.0020	-0.0080	-0.0050	-0.0025
β_{TC}	-0.0280	-0.0140	-0.0140	-0.0560	-0.0100	-0.0050
β_{AT}	-0.0080	-0.0040	-0.0040	-0.0160	-0.0120	-0.0060
α_{TT_R}	1.2500	1.2500	1.2500	1.2500	0.8000	0.8000
α_{TT_A}	2.0000	2.0000	2.0000	2.0000	1.7000	1.7000
α_{TT_H}	1.5000	1.5000	1.5000	1.5000	1.7000	1.7000
Share(Car)	53.06%	54.74%	54.64%	46.30%	43.64%	53.52%
Share(Rail)	9.54%	5.52%	7.80%	9.68%	6.22%	5.44%
Share(Air)	13.00%	10.58%	12.20%	16.84%	39.08%	26.34%
Share(HSR)	24.40%	29.16%	25.36%	27.18%	11.06%	14.70%

3.2 Results from simulated datasets

For each of the simulated datasets, we estimate 14 different models. The first four of these are basic RUM, RRM, P-RRM and RAM models. The next 10 are latent class models with two classes, using all possible combinations of models, i.e. RUM_RUM, RUM_P-RRM, etc. This means that for each dataset, we have four models that test for taste heterogeneity alone and six models that allow for taste and decision rule heterogeneity. As we have 11 different datasets that include taste and/or decision rule heterogeneity, we aim to test the findings obtained using simultaneous LC models and sequential LC models, i.e. model averaging, and contrast these with the true DGP. The results of all models are given in Table 3. We highlight the best-fitting base model for each DGP, together with base models that receive a share from model averaging across these models. The table also details the best-fitting latent class model and the gain achieved by averaging across the latent class models.

Table 2 : Types of heterogeneity in each dataset.

Dataset	Data Generation Process				Sources of heterogeneity	
					Taste	Decision rule
1	RUM1 (100%)				no	no
2	P-RRM1 (100%)				no	no
3	RUM1 (50%)	RUM2 (50%)			yes	no
4	P-RRM1 (50%)	P-RRM2 (50%)			yes	no
5	RUM1 (50%)	P-RRM1 (50%)			no	yes
6	RUM2 (50%)	P-RRM2 (50%)			no	yes
7	RUM1 (25%)	P-RRM1 (25%)	RRM1 (25%)	RAM1 (25%)	no	yes
8	RUM1 (50%)	P-RRM2 (50%)			yes	yes
9	RUM2 (50%)	P-RRM1 (50%)			yes	yes
10	RUM1 (50%)	RUM2 (50%)	P-RRM1 (25%)	P-RRM2 (25%)	yes	yes
11	RUM2 (25%)	P-RRM2 (25%)	RRM1 (25%)	RAM1 (25%)	yes	yes

We first look at the two datasets which come from a single model without heterogeneity (i.e. datasets 1-2). We note that each time, the model using the correct decision rule outperforms the other models as well as obtaining large shares from model averaging across the base models. None of the LC structures can reject the single class model, with a maximum gain of 15.94 units at a cost of 14 additional parameters, resulting in Likelihood Ratio Test p-values of 0.64 and 0.32 respectively for datasets 1 and 2. Consequently we do not detail results from model averaging across latent class models for these datasets given that these models are rejected.

As a contrast, we observe substantial gains in model fit from moving to latent class models for datasets 3 and 4, with highly significant likelihood ratio tests. In both cases, one model (RUM_RUM and P-RRM_P-RRM, respectively) obtains a large share from model averaging across the latent class models. This gives a good indication of the underlying DGP, which are correctly identified in both of these cases.

For datasets 5-7, which are generated with decision-rule heterogeneity alone, we observe more similar log-likelihoods for the base models, with the shares from model averaging indicating the DGP. For the first two cases, the power of model averaging is particularly highlighted given that RRM is the best performing base model, but receives none of the model averaging share. Unsurprisingly, the best performing latent class model for both datasets are the RUM_P-RRM models, but these models are again rejected as was the case for datasets 1 and 2, as the improvement in model fit is not substantial given the increase in model parameters and consequently we only have weak evidence from the likelihood ratio tests. For dataset 7, it is notable that model averaging gives a share to all base models. Though the gain from model averaging across base models is less than the gain from moving to latent class models (as has to be the case), these results indicate that decision rule heterogeneity alone is present in datasets 5-7. This is indicated by the fact that averaging over the base models achieves at least 20% of the gain that is achieved by a latent class model. As a contrast, this gain is no higher than 3% when the dataset additionally includes taste heterogeneity. Whilst the rejection of the latent class models could be inferred as

Table 3 : Results from RUM, RRM, P-RRM, RAM and latent class models for each of the 11 simulated datasets, together with the results from model averaging.

DGP Type of heterogeneity		None		Taste		Decision-rule (DR)	
Dataset		1	2	3	4	5	6
Data Generation Process (DGP)		MNL1 (100%)	PRRM1 (100%)	MNL1 (50%) MNL2 (50%)	PRRM1 (50%) PRRM2 (50%)	MNL1 (50%) PRRM1 (50%)	MNL2 (50%) PRRM2 (50%)
Base	MNL	-4,816.02	-4,403.37	-5,293.13	-4,923.14	-4,556.99	-5,112.98
models	RRM	-4,841.62	-4,360.38	-5,304.37	-4,910.58	-4,551.88	-5,112.19
LL(0) =	PRRM	-4,909.46	-4,340.02	-5,346.80	-4,903.64	-4,570.67	-5,116.47
-6,931.47	RAM	-5,025.86	-4,468.43	-5,376.25	-5,001.32	-4,764.95	-5,186.00
Model Averaging (base models) LL		-4,816.02	-4,339.50	-5,292.99	-4,901.05	-4,543.79	-5,107.61
Model averaging LL - best LL (Base models)		0.00	0.52	0.15	2.59	8.09	4.58
	MNL	100.00%	0.00%	96.15%	14.16%	57.66%	55.58%
Base models	RRM	0.00%	3.96%	0.00%	13.28%	0.00%	0.00%
shares	PRRM	0.00%	91.98%	0.00%	70.33%	39.37%	44.42%
	RAM	0.00%	4.06%	3.85%	2.24%	2.97%	0.00%
Best LL (latent class model)		MNL-MNL	MNL-PRRM	MNL-MNL	PRRM-PRRM	MNL-PRRM	MNL-PRRM
Log-likelihood		-4,804.49	-4,324.07	-5,082.81	-4,758.47	-4,529.72	-5,092.82
Best LL (LC models) - best LL (base model)		11.53	15.94	210.32	145.17	22.15	19.37
Likelihood Ratio Test (p-value)		0.6438	0.3168	0.0000	0.0000	0.0755	0.1513
Model averaging LL (over LC models)		n/a		-5,082.18	-4,756.46	n/a	
Model averaging LL - best LL (LC models)				0.64	2.01		
Best LL (LC model) model averaging share				76.55%	82.38%		

DGP Type of heterogeneity		DR	Taste & DR			
Dataset		7	8	9	10	11
Data Generation Process (DGP)		MNL1 (25%) PRRM1 (25%) RRM1 (25%) RAM1 (25%)	MNL1 (50%) PRRM2 (50%)	MNL2 (50%) PRRM1 (50%)	MNL1 (25%) MNL2 (25%) PRRM1 (25%) PRRM2 (25%)	MNL2 (25%) PRRM2 (25%) RRM1 (25%) RAM1 (25%)
Base	MNL	-4,671.91	-5,114.66	-5,224.66	-5,118.18	-5,223.75
models	RRM	-4,667.38	-5,114.63	-5,226.79	-5,124.55	-5,227.62
LL(0) =	PRRM	-4,700.63	-5,136.53	-5,246.89	-5,156.06	-5,257.52
-6,931.47	RAM	-4,711.57	-5,229.82	-5,287.68	-5,196.98	-5,281.99
Model Averaging (base models) LL		-4,657.39	-5,112.00	-5,222.46	-5,117.78	-5,222.17
Model averaging LL - best LL (Base models)		9.99	2.63	2.21	0.40	1.58
	MNL	24.13%	71.76%	76.58%	90.95%	72.55%
Base models	RRM	34.79%	6.45%	0.00%	0.00%	7.80%
shares	PRRM	14.62%	21.79%	22.27%	9.05%	8.92%
	RAM	26.46%	0.00%	1.15%	0.00%	10.73%
Best LL (latent class model)		RRM-RAM	MNL-PRRM	MNL-PRRM	RRM-RRM	RRM-RRM
Log-likelihood		-4,644.55	-5,017.54	-4,927.19	-4,936.38	-5,063.79
Best LL (LC models) - best LL (base model)		22.83	97.09	297.48	181.80	159.95
Likelihood Ratio Test (p-value)		0.0630	0.0000	0.0000	0.0000	0.0000
Model averaging LL (over LC models)		n/a	-5,016.83	-4,925.71	-4,933.23	-5,054.26
Model averaging LL - best LL (LC models)			0.71	1.48	3.15	9.53
Best LL (LC model) model averaging share			72.98%	76.46%	21.27%	29.61%

'failing to uncover' the DGP, it actually points to the fact that the parameter values across decision-rules are similar enough that taste heterogeneity does not exist.

Finally, for the datasets created with both types of heterogeneity (where the models of different types also use substantially different relative taste coefficients), we observe substantial improvements in model fit by moving to latent class models, as expected. Model averaging across these latent class models then tells us if the best-fitting latent class model is the DGP for the dataset, with RUM_P-RRM obtaining 73% and 76% of the shares for datasets 8 and 9 respectively, but RRM_RRM only obtaining 21% and 30% for datasets 10 and 11. These latter cases imply that there is more than two processes in the underlying DGP for these datasets.

4. Analysis on SP data

This section presents our work on two typical SP datasets. We first give an overview of the data before looking separately at the case of attribute non-attendance (cf. Section 4.2) and decision rule heterogeneity (cf. Section 4.3).

4.1 Data

Our main analysis relies on two SP datasets. The first is from Hess and Stathopoulos (2013) and the second is developed by Swärdh and Algers (2009), with descriptions also in Beck and Hess (2016).

SP dataset 1

For the first dataset, public transport commuters living in the UK each make ten choices between three routes. A total of 368 participants completed the survey resulting in 3,680 choices. Each choice task involves an invariant reference trip and two hypothetical alternatives (where the invariant trip is chosen 35.19% of the time and the new alternatives have shares of 34.27% and 30.54%, respectively). The invariant trip for each individual is based on averaging trip attributes across 10 regular trips corresponding to a week of commuting, with the attributes of the hypothetical alternatives being pivoted around those of the invariant trip. These choice tasks were generated with a D-efficient experimental design using NGene. A total of 60 choice scenarios were blocked into groups of 10. Further details for the dataset are given by Hess and Stathopoulos (2013). Each alternative is described by travel time (in minutes), fare (in £), rate of crowded trips, rate of delays (both out of 10 trips), the average length of delays (across delayed trips) and the presence of a delay information service (either not available, available at a small fixed cost, or free). This dataset has previously been used for decision rule heterogeneity (Hess and Stathopoulos, 2013) as well as for ANA work (Hess et al., 2013b), making it an ideal case study for the present paper.

SP dataset 2

The second dataset used in this work involves decision-makers completing two distinct sets of choice tasks based on an individual's willingness to accept longer commutes for better salaries (see Beck and Hess, 2016, for a detailed description of the survey). A sample of 1,179 households (with both partners in each household, resulting in 2,358 individuals) completed 4 tasks involving only attributes affecting themselves, and 4 or 5 tasks with attributes impacting both members of the household. This resulted in a total of 20,041 choice observations. All choice tasks included trade-offs between the individual's current travel time and salary or an increased salary (of 500 or 1000 Swedish Krona (SEK) in net wage per month) at a cost of an increase in one-way travel time (of either 10 or 25 minutes). Similar adjustments were also made to the salary and travel time of the partner in choice tasks also affecting the partner. All choice tasks included a status quo alternative, a new location and an 'I am indifferent' option. This dataset is also well suited to for exploring different sources of heterogeneity. Whilst it has previously been used to demonstrate that random regret minimisation is more suited than random utility models at capturing choice indifference Hess et al. (2014), 79% of individuals never choose the indifference alternative. This could result in the presence of decision rule heterogeneity as there is no reason to assume that RRM models will best fit these individuals also. Furthermore, there is scope for attribute non-attendance as some individuals may focus on travel time or salary only or alternatively may consider attributes affecting themselves but not their partners.

4.2 Attribute non-attendance work

We first look at the case of ANA, where we adopt a specification in line with Hess et al. (2013b).

Specification

We start by estimating a simple RUM model. For SP-1, we use a logarithmic transform on the fare attribute given earlier evidence of strong non-linearity. The model uses five marginal utility parameters for the continuous attributes, two parameters for the dummy coded delay information system, and two alternative specific constants (ASC). For SP-2, the model uses different sets of parameters for the choice tasks involving attributes impacting just the decision-maker and those with attributes also impacting the partner. This results in six marginal utility parameters and four alternative specific constants. The model follows Hess et al. (2014) in setting the utility for the indifference alternative to a constant.

We next move to the latent class model for attribute non-attendance. We use models with 2^K classes, with all combinations of attendance and non-attendance for the K parameters. The probability for class s is given by π_s, with $0 \leq \pi_s \leq 1$ and $\sum_{s=1}^{S} \pi_s = 1$. Rather than imposing constraints in estimation, an easier approach is to use $\pi_s = \frac{e^{\delta_s}}{\sum_{m=1}^{S} e^{\delta_m}}$, with one δ_m, i.e. the parameter used in the class allocation probabilities, being fixed to zero. Nevertheless, this specification still involves estimating $2^K - 1$ separate δ terms, of which many will be very negative, equating to very small class probabilities. In the context of the applications presented in this paper, we make the simplifying assumption that attendance versus non-attendance is independent across attributes (with probabilities that vary across attributes but are constant across individuals), by instead setting

$$\pi_s = \prod_{k=1}^{K} \left(\Lambda_{s,k} \left(1 - P_{\text{ANA},k}\right) + \left(1 - \Lambda_{s,k}\right) P_{\text{ANA},k} \right), \tag{14}$$

where $\Lambda_{s,k}$ gives the entry in Λ relating to attribute k in class s, where this is 1 only if attribute k is attended to in class s. With this specification, we only need to estimate K separate δ elements (as $P_{\text{ANA},k}$ is the probability of non-attendance to attribute k, thus $P_{\text{ANA},k} = \frac{e^{\delta_k}}{e^{\delta_k}+1}$), as opposed to $2^K - 1$, leading to significant reductions in the number of parameters.

We finally look at the estimation of our model averaging structure. For this, we first estimate 128 and 64 individual models for the two datasets respectively, corresponding to all possible combinations of attribute attendance and non-attendance, i.e. for SP-1, going from a model with all 9 model parameters (all 7 attributes are attended to) to one with the two alternative specific constants only (none of the attributes attended to). We then estimate the model averaging structure, meaning that we keep the parameters for each of the 128/64 models at the estimates from the individual model estimation process and only estimate the weights for model averaging. We again use multiplicative class allocation probabilities, as in the LC model.

Results for SP-1

The results for the simple RUM model are shown in Table 4 where all estimates are of the expected sign.

Table 4 : RUM results for SP-1.

LL(0)	-4,042.89
LL(final)	-3,366.95
ρ^2	0.1672
adj. ρ^2	0.1650

	Estimate	Rob.t.ratio(0)
ASC_1	0.3841	5.76
ASC_2	0.1608	3.26
$\beta_{\text{travel time}}$	-0.0467	-9.47
$\beta_{\text{log-fare}}$	-5.9726	-18.89
β_{crowding}	-0.2198	-8.51
$\beta_{\text{rate of delays}}$	-0.2411	-9.82
$\beta_{\text{average delay}}$	-0.0421	-5.35
$\beta_{\text{info system charged}}$	-0.0833	-1.04
$\beta_{\text{info system free}}$	0.3370	5.06

The results for the CLC model are shown in Table 5. We see an improvement in log-likelihood by 308.16 units for 7 additional parameters. This is highly significant and in line with previous findings when using such a CLC model for ANA. We also see that the magnitudes of the marginal utility parameters, which now only apply in the attendance classes, have increased substantially compared to the base model. This is what we expect as the RUM model has to find a single value to represent the importance of the attribute to all decision-makers, which is between 0 (for the non-attenders) and the observed estimate (for the attenders) from the confirmatory latent class model. The exception is for $\beta_{\text{info system charged}}$, which has an insignificant negative coefficient in the RUM model, but becomes significantly positive under the new model. This is a result of the attribute having a very high rate of non-attendance (96%), which causes it to be insignificant at the group level. The implied rates of non-attendance of all attributes are in fact unrealistically high, exceeding 50% for all attributes except fare.

Table 5 : Confirmatory latent class model for attribute non-attendance for SP-1, with model estimates including implied rates of attribute non-attendance (ANA).

LL(0)	-4,042.89	
LL(final)	-3,058.79	
ρ^2	0.2434	
adj. ρ^2	0.2395	
	Estimate	Rob.t.ratio(0)
ASC_1	0.8416	10.32
ASC_2	0.3290	4.23
$\beta_{\text{travel time}}$	-0.1841	-5.64
$\beta_{\text{log-fare}}$	-14.6889	-14.37
β_{crowding}	-1.1524	-7.16
$\beta_{\text{rate of delays}}$	-1.1307	-5.62
$\beta_{\text{average delay}}$	-0.3966	-4.85
$\beta_{\text{info system charged}}$	2.3264	3.37
$\beta_{\text{info system free}}$	2.0433	7.23
$\delta_{\text{ANA,travel time}}$	0.3232	1.11
$\delta_{\text{ANA,log-fare}}$	-0.5142	-3.43
$\delta_{\text{ANA,crowding}}$	0.7767	3.30
$\delta_{\text{ANA,rate of delays}}$	0.7363	2.43
$\delta_{\text{ANA,average delay}}$	1.1917	4.02
$\delta_{\text{ANA,info system charged}}$	3.1776	3.82
$\delta_{\text{ANA,info system free}}$	0.9874	3.61
	Implied rate of ANA	
	Estimate	Rob.t.ratio(0)
travel time	0.5801	8.18
fare	0.3742	10.65
crowding	0.6850	13.49
rate of delays	0.6762	10.21
average delay	0.7670	14.48
info system charged	0.9600	30.05
info system free	0.7286	13.47

For model averaging, we initially estimated seven class allocation weights as in the LC model but find that for the first four attributes, the constants go towards $-\infty$, suggesting a zero probability of ANA. The results of the model averaging work are shown in Table 6. We see that this model now only offers a marginally better log-likelihood than the RUM model in Table 4, much in contrast with the LC model in Table 5. No formal statistical test is used here as model averaging is not a process of simultaneously estimating all the parameters for all the models on a single dataset. In addition to the earlier finding of zero weight for any classes that imply non-attendance of either time, fare, crowding or the rate of delays, we also see low rates for the average delay and the free information system, with a higher rate for the charged system. A number of other statistics are valuable. First, we can rank the 128 models by log-likelihood and we note that the 8 models that obtain the best individual

Table 6 : Model averaging (MA) for ANA work for SP-1.

LL(final) = -3,363.28, LL(0) = -4,042.89			Implied rate of ANA		
	Estimate	Rob.t.ratio(0)		Estimate	Rob.t.ratio(0)
$\delta_{\text{ANA,average delay}}$	-1.9099	-1.95	average delay	0.1290	1.17
$\delta_{\text{ANA,ch inf sys}}$	0.0844	0.05	info system charged	0.5211	1.22
$\delta_{\text{ANA,free inf sys}}$	-1.1531	-2.04	info system free	0.2399	2.33

	Information for 8 retained models.							
LL	-3,367.75	-3,366.95	-3,400.98	-3,390.17	-3,391.85	-3,391.62	-3,424.48	-3,416.22
ranking out of 128 candidates	2	1	6	3	5	4	8	7
providing best fit for N respondents	12	17	14	14	9	8	12	9
MA share	34.50%	31.71%	10.89%	10.01%	5.11%	4.70%	1.61%	1.48%
	attribute included							
travel time	YES	YES	YES	YES	YES	YES	YES	YES
fare	YES	YES	YES	YES	YES	YES	YES	YES
crowding	YES	YES	YES	YES	YES	YES	YES	YES
rate of delays	YES	YES	YES	YES	YES	YES	YES	YES
average delay	YES	YES	YES	YES	NO	NO	NO	NO
info system charged	NO	YES	NO	YES	NO	YES	NO	YES
info system free	YES	YES	NO	NO	YES	YES	NO	NO
	est. (rob. t-rat)							
ASC_1	0.41 (6.46)	0.38 (5.76)	0.40 (6.24)	0.32 (4.77)	0.39 (6.15)	0.38 (5.61)	0.38 (5.91)	0.31 (4.61)
ASC_2	0.16 (3.29)	0.16 (3.26)	0.16 (3.29)	0.16 (3.17)	0.18 (3.59)	0.18 (3.58)	0.17 (3.46)	0.17 (3.41)
$\beta_{\text{travel time}}$	-0.05 (-9.48)	-0.05 (-9.47)	-0.05 (-9.73)	-0.05 (-9.63)	-0.05 (-9.35)	-0.05 (-9.34)	-0.05 (-9.59)	-0.05 (-9.49)
$\beta_{\text{log-fare}}$	-5.95 (-18.86)	-5.97 (-18.89)	-5.77 (-18.19)	-5.90 (-18.62)	-5.87 (-18.81)	-5.88 (-18.81)	-5.68 (-18.12)	-5.80 (-18.51)
β_{crowding}	-0.22 (-8.50)	-0.22 (-8.51)	-0.22 (-8.59)	-0.22 (-8.61)	-0.22 (-8.46)	-0.22 (-8.46)	-0.22 (-8.55)	-0.22 (-8.56)
$\beta_{\text{rate of delays}}$	-0.24 (-9.76)	-0.24 (-9.82)	-0.24 (-9.82)	-0.24 (-9.95)	-0.27 (-10.94)	-0.27 (-10.98)	-0.26 (-11.00)	-0.27 (-11.17)
$\beta_{\text{average delay}}$	-0.04 (-5.32)	-0.04 (-5.35)	-0.04 (-5.29)	-0.04 (-5.51)	0	0	0	0
$\beta_{\text{info system charged}}$	0	-0.08 (-1.04)	0	-0.27 (-3.67)	0	-0.04 (-0.57)	0	-0.24 (-3.24)
$\beta_{\text{info system free}}$	0.36 (5.96)	0.34 (5.06)	0	0	0.36 (5.91)	0.35 (5.22)	0	0

log-likelihoods are also the only 8 models that contribute to the model average. The two best fitting models also contribute the most to the model averaging, though in reverse order (models with rankings 2 and 1). Finally, for each individual person in the data, we can see which of the 128 models best explains their choices. Doing this, we see that out of the 368 individuals in the data, only 95 have their choices explained the best way by one of these 8 models, where a remarkable 104 out of the 128 models have at least one individual where they are the best performing model.

Overall, the findings from this analysis are much in contrast with those from the confirmatory latent class model in that very little evidence of ANA is found. In addition, there is very little variation in the remaining parameters across classes. Of course, the counter-argument could be that the model averaging approach cannot retrieve ANA as it is based on individual models that each apply a homogeneous approach to all individuals. However, some reassurance can be obtained from the fact that the model averaging results are in line with the findings by Hess et al. (2013b) which find evidence of ANA only for the average delay attribute and for the delay information attribute after allowing for random heterogeneity in their models. It is thus doubtful whether additional insights would be obtained with more flexibility for the individual models, such as by including random heterogeneity. A possible step in that direction would be to estimate one latent class model for each of the 128 candidates, i.e. allowing for heterogeneity within a model that assumes a given ANA strategy.

Results for SP-2

The results for the simple RUM model are shown in Table 7 where all estimates are of the expected sign. Decision-makers give more weight to their own salary than their partner's salary but the reverse is true for travel time.

Table 7 : RUM results for SP-2.

LL(0)		-22,017.29	
LL(final)		-14,153.13	
ρ^2		0.3572	
adj. ρ^2		0.3567	
		Estimate	Rob.t.ratio(0)
First set of choice tasks	ASC_{base}	0.5039	10.09
	$ASC_{\text{indifference}}$	-2.904	-11.96
	$\beta_{\text{own-travel-time}}$	-0.0335	-14.00
	$\beta_{\text{own-salary}}$	0.0136	2.38
Second set of choice tasks	ASC_{base}	0.8878	14.63
	$ASC_{\text{indifference}}$	-1.8309	-5.44
	$\beta_{\text{own-travel-time}}$	-0.0129	-5.93
	$\beta_{\text{own-salary}}$	0.0178	3.21
	$\beta_{\text{partner-travel-time}}$	-0.0145	-6.88
	$\beta_{\text{partner-salary}}$	0.0118	1.95

We next consider a confirmatory latent class model, with the results given in Table 8. We again see a substantial improvement in model fit, this time of 1,696 units. The magnitude of the marginal utility parameters increase in comparison to the base model, with a decision-maker's own salary in particular becoming substantially more important than the partner's salary. As a contrast to the results of SP-1, the implied rates of attribute non-attendance appear more reasonable, with salaries more often ignored than travel time.

Table 8 : Confirmatory latent class model for attribute non-attendance for SP-2, with model estimates including implied rates of attribute non-attendance.

LL(0)		-22,017.29	
LL(final)		-12,457.15	
ρ^2		0.4342	
adj. ρ^2		0.4335	
		Estimate	Rob.t.ratio(0)
First set of choice tasks	ASC_{base}	-0.2674	-3.72
	$ASC_{\text{indifference}}$	-3.9904	-14.43
	$\beta_{\text{own-travel-time}}$	-0.1080	-21.27
	$\beta_{\text{own-salary}}$	0.4152	8.02
	$\delta_{\text{ANA,own-travel-time}}$	-2.8640	-20.23
	$\delta_{\text{ANA,own-salary}}$	-1.2640	-14.96
Second set of choice tasks	ASC_{base}	-1.3283	-8.02
	$ASC_{\text{indifference}}$	-5.1077	-10.18
	$\beta_{\text{own-travel-time}}$	-0.1452	-25.18
	$\beta_{\text{own-salary}}$	1.6070	7.68
	$\beta_{\text{partner-travel-time}}$	-0.1823	-26.50
	$\beta_{\text{partner-salary}}$	0.2019	12.81
	$\delta_{\text{ANA,own-travel-time}}$	-1.9540	-15.56
	$\delta_{\text{ANA,own-salary}}$	-1.4494	-18.55
	$\delta_{\text{ANA,partner-travel-time}}$	-1.7783	-18.13
	$\delta_{\text{ANA,partner-salary}}$	-0.9168	-5.96
		Implied rate of ANA	
		Estimate	Rob.t.ratio(0)
first own-travel-time		0.0540	7.47
first own-salary		0.2203	15.18
second own-travel-time		0.1241	9.09
second own-salary		0.1901	15.80
partner-travel-time		0.1445	11.91
partner-salary		0.2856	9.10

For model averaging, we estimate six class allocation weights as in the LC model. We find that the constants for the decision-maker's own travel time go towards $-\infty$, suggesting a zero probability of ANA. The results of model averaging are displayed in Table 9. 15 different models contribute to the model average, with the four best performing models also being the four with the largest shares in model averaging. 61 out of 64 of the models are the best performing model for at least one individual. As was the case for SP-1, we observe a substantial reduction in the improvement offered over the base model, with a gain of just 12 units instead of 1,696. However, as a contrast, the implied rates of attribute non-attendance are more in line with those of the CLC model, with the rates for attributes impacting the partner particularly similar.

Table 9 : Model averaging for ANA work for SP-2.

LL(final) = -14,140.92, LL(0) = -22,017.92			Implied rate of ANA		
	Estimate	Rob.t.ratio(0)		Estimate	Rob.t.ratio(0)
$\delta_{\text{NA,first-own-salary}}$	-1.0459	-1.44	first-own-salary	0.2600	1.85
$\delta_{\text{NA,second-own-salary}}$	-1.0273	-1.75	second own-salary	0.2636	2.32
$\delta_{\text{NA,partner-travel-time}}$	-1.692	-2.81	partner-travel-time	0.1555	1.97
$\delta_{\text{NA,partner-salary}}$	-0.7669	-1.21	partner-salary	0.3172	2.30

	Information for top 4 retained models			
individual LL	-14,153.13	-14,160.26	-14,446.75	-14,474.33
ranking out of 64 models	1	2	4	3
providing best fit for N respondents	115	183	30	11
MA share	31.42%	14.59%	11.25%	11.04%
	attribute included			
first own-travel-time	YES	YES	YES	YES
first own-salary	YES	YES	YES	NO
second own-travel-time	YES	YES	YES	YES
second own-salary	YES	YES	NO	YES
partner-travel-time	YES	YES	YES	YES
partner-salary	YES	NO	YES	YES
	est. (rob. t-rat)			
ASC_{base}	0.504 (10.09)	0.504 (10.09)	0.504 (10.09)	0.493 (9.81)
$ASC_{\text{indifference}}$	-2.904 (-11.96)	0.884 (14.51)	0.879 (14.52)	0.888 (14.63)
$\beta_{\text{own-travel-time}}$	0.034 (-14.00)	-0.034 (-14.00)	-0.034 (-14.00)	-0.034 (-13.83)
$\beta_{\text{own-salary}}$	0.014 (2.38)	0.014 (2.39)	0.014 (2.39)	0
ASC_{base}	0.888 (14.63)	-2.904 (-11.95)	-2.904 (-11.95)	-3.319 (-18.27)
$ASC_{\text{indifference}}$	-1.831 (-5.44)	-2.118 (-7.64)	-2.271 (-7.75)	-1.831 (-5.44)
$\beta_{\text{own-travel-time}}$	-0.013 (-5.93)	-0.013 (-5.90)	-0.013 (-5.79)	-0.013 (-5.93)
$\beta_{\text{own-salary}}$	0.018 (3.21)	0.02 (3.41)	0	0.018 (3.21)
$\beta_{\text{partner-travel-time}}$	-0.015 (-6.88)	-0.014 (-6.72)	-0.015 (-7.01)	-0.015 (-6.88)
$\beta_{\text{partner-salary}}$	0.012 (1.95)	0	0.015 (2.33)	0.012 (1.95)

4.3 Decision rule heterogeneity work

We next turn to decision rule heterogeneity, which has been the key interest in applying latent class structures for process heterogeneity in recent years. We use the same four decision rules (RUM, RRM, P-RRM and RAM) that we used in our simulated datasets.

Results for SP-1

For SP-1, we first apply the four different models individually, obtaining the results given in Table 10. We see that RAM obtains the best log-likelihood and Bayesian Information Criterion (BIC) ahead of RUM, while the performance of the two regret-based models is comparatively worse. As a first step, we look at model averaging across these four individual models with different decision rules, where the resulting shares and fit are shown in Table 10. We see that the model average leads to no improvement in model fit, as RAM is given 100% of the share.

Table 10 : Results from different individual models applied to the SP-1 dataset.

Model	Model Type	Log-likelihood	BIC	MA Share
1	RUM	-3,366.95	6,808	0.00%
2	RRM	-3,371.01	6,816	0.00%
3	P-RRM	-3,407.72	6,889	0.00%
4	RAM	-3,361.29	6,796	100.00%
Model averaging		-3,361.29		

In practice, the estimation of a latent class model with four separate classes all using individual decision rules is computationally challenging and most applications rely on combining just two different rules. We therefore look at the estimation of 10 different latent class structures with two classes each, picking all combinations of two model structures with replacement, thus also allowing for four models where the two classes are of the same type, i.e. looking for taste heterogeneity alone. Table 11 gives the log-likelihoods of these models. For all 10 models, a likelihood ratio test against the corresponding model (in the case of single decision rule) or two corresponding models (in the case of two decision rules) clearly rejects the base model(s). This provides evidence of taste heterogeneity (in the case of single structure models) and would typically be seen as evidence of decision rule heterogeneity in the case of the models with two different structures in the two classes.

Table 11 : Results from latent class models applied to the SP-1 dataset.

Model version	Class 1	Class 2	Log-likelihood	BIC	MA Share
1	RUM	RUM	-3,118.28	6,393	0.0%
2	RUM	RRM	-3,107.12	6,370	4.4%
3	RUM	P-RRM	-3,115.52	6,387	0.0%
4	RUM	RAM	-3,117.41	6,391	0.0%
5	RRM	RRM	-3,111.32	6,379	0.0%
6	RRM	P-RRM	-3,136.37	6,429	0.0%
7	RRM	RAM	-3,106.33	6,369	27.1%
8	P-RRM	P-RRM	-3,146.02	6,448	0.0%
9	P-RRM	RAM	-3,114.72	6,385	0.0%
10	RAM	RAM	-3,105.15	6,366	59.2%
Best LL (LC models) - best LL (base model)			256.14		
Model averaging			-3,100.72		
Gain from model averaging			4.43		

Most existing applications compare a model combining multiple different decision rules to a set of single class models using the individual rules. This comparison is of course likely to be biased in the presence of taste heterogeneity. Crucially, the improvements to be made from combining different structures depend on their individual performance. For example, we see that for RAM, which is the best performing individual model in Table 10, combining the model with a different structure does not reach as high a log-likelihood as a structure with two separate RAM classes. On the other hand, for those models that perform less well individually, combining them with a different structure gives a better log-likelihood than a model with two classes using the same structure. This already suggests that the results from the latent class structure point more towards taste heterogeneity than decision rule heterogeneity. In fact, when looking at pairs of decision rules, we see only two cases in favour of decision rule heterogeneity, i.e. where a model combining two decision rules outperforms the two LC models that use the same model type in both classes. The RUM_RRM model outperforms RRM_RRM by 4.20 log-likelihood units and outperforms RUM_RUM by 11.16 units. Additionally, RUM_P-RRM has a better log-likelihood than either RUM_RUM or P-RRM_P-RRM. Further evidence is given in the model averaging results in Table 11, with 59% of the share going to a single model. In comparison to the outputs from our simulated datasets, these results are most similar to the cases with taste heterogeneity alone. The gain from averaging across these latent class models is however 4.43 units, which is slightly more than observed for datasets with taste heterogeneity alone. Overall, our findings highlight the importance of within-model taste heterogeneity.

To examine this further, we explore the most common example of decision rule heterogeneity (RUM-RRM) in more detail by also considering the outputs for the parameter estimates, in comparison to a model average performed on RUM and RRM. The results for this are shown in Table 12.

Table 12 : A detailed example of model averaging compared to a simultaneous latent class approach using RUM and RRM.

	Latent Class - 1 model		Model averaging - 3 models	
	19 pars, estimated simultaneously		2*9 pars, then 1 for MA	
	Class 1:RUM	Class 2:RRM	Class 1: RUM	Class 2: RRM
Class LL:	-3,643.18	-4,503.18	-3,366.95	-3,371.01
Log-likelihood	-3,107.12		-3,366.94	
ASC_1	0.63 (6.44)	0.05 (0.39)	0.38 (5.76)	0.26 (4.03)
ASC_2	0.24 (2.91)	0.20 (1.26)	0.16 (3.26)	0.16 (3.33)
$\beta_{\text{travel time}}$	-0.05 (-6.85)	-0.06 (-6.85)	-0.05 (-9.47)	-0.03 (-9.57)
$\beta_{\text{log-fare}}$	-3.22 (-7.17)	-11.61 (-7.65)	-5.97 (-18.89)	-4.08 (-17.70)
β_{crowding}	-0.31 (-7.14)	-0.15 (-2.70)	-0.22 (-8.51)	-0.14 (-8.53)
$\beta_{\text{rate of delays}}$	-0.39 (-4.20)	-0.01 (-0.44)	-0.24 (-9.82)	-0.16 (-9.93)
$\beta_{\text{average delay}}$	-0.06 (-8.79)	-0.12 (-3.04)	-0.04 (-5.35)	-0.03 (-5.16)
$\beta_{\text{info system charged}}$	-0.09 (-0.79)	0.00 (-1.13)	-0.08 (-1.04)	-0.05 (-0.86)
$\beta_{\text{info system free}}$	0.54 (5.97)	0.55 (0.55)	0.33 (5.06)	0.22 (4.97)
π_m	60.10% (2.11)	39.90%	93.02% (2.77)	6.98%

Table 12 gives model fit as well as estimates for the above parameters for both a latent class model and a model averaging approach. The model averaging approach separately runs RUM and RRM models before then estimating a class allocation parameter individually. Crucially, the model averaging approach does not result in a significant improvement over a RUM model on its own, with an improvement of just 0.01 log-likelihood units. As a contrast, the latent class approach results in a vast improvement in model fit (260 units). At face value, this would again suggest decision rule heterogeneity, although the fit is not much better than for the RUM-RUM or RRM-RRM models. Most significantly, it appears that the fare parameter estimates (highlighted in red) are very different between the two classes. In contrast with the model averaging results, and given the poor class-specific model fit for the RRM class (compared to the RRM-RRM model), we believe that this finding shows that a substantial share of the improvements obtained by the RUM-RRM model are due to heterogeneity in the cost sensitivity rather than heterogeneity in the decision rules. This means that the classes individually have a very poor fit (as they cannot explain all individuals) but when combined into a latent class approach, the result is a model with far superior model fit. Together with the poor improvement from model averaging, these results suggest that most of the model improvement for this dataset is due to taste rather than decision rule heterogeneity.

Results for SP-2

For SP-2, we first apply the four different models individually, obtaining the results given in Table 13. In line with the results of Hess et al. (2014), we observe that RRM gives the best model fit. P-RRM has slightly worse fit but both of these models have substantially better model fit than RUM and RAM, both of which appear to be unable to capture the presence of an indifference alternative. However, despite the vastly inferior model fit, RUM still obtains a 5% share from model averaging. Additionally, the shares are not necessarily proportional to the model fit of the individual model, as P-RRM obtains a larger share than RRM, despite having poorer overall individual log-likelihood. This again shows that some models can work well for some decision-makers even if they obtain a lower overall fit to the sample. Overall, the model average results in an improvement of 117 units over the RRM model, implying that there is likely decision rule heterogeneity in this case.

Table 13 : Results from different individual models applied to the SP-2 dataset.

Model	Model Type	Log-likelihood	BIC	MA Share
1	RUM	-14,153.13	28,388	5.42%
2	RRM	-12,426.76	24,936	42.12%
3	P-RRM	-12,438.01	24,958	52.46%
4	RAM	-14,423.92	28,930	0.00%
Model averaging		-12,309.80		
Gain from model averaging		116.96		

We next look at the estimation of the 10 different latent class models with two classes each. Table 14 gives the log-likelihood of these models.

In all cases, we observe a substantial improvement in log-likelihood, with all 10 models having a better BIC than that of the best individual model. Overall, the best latent class model is the P-RRM_P-RRM model, which records a likelihood that is 1,706 units better than the RRM model from Table 13. Notably, the best performing latent class model for SP-1 (RAM-RAM) is the worst performing model for SP-2. In this case, a lack of decision rule heterogeneity is implied by the fact that not a single model combining two decision rules outperforms the two LC models that use the same model type in both classes. However, averaging across all 10 of the latent class models results in a 410 unit improvement in model fit. With the best latent class model only receiving 33% of the model averaging share,

Table 14 : Results from latent class models applied to the SP-2 dataset.

Model version	Class 1	Class 2	Log-likelihood	BIC	MA Share
1	RUM	RUM	-12,093.16	24,359	2.7%
2	RUM	RRM	-11,249.03	22,670	0.0%
3	RUM	P-RRM	-11,281.73	22,736	32.5%
4	RUM	RAM	-12,098.96	24,370	0.0%
5	RRM	RRM	-10,937.65	22,048	4.5%
6	RRM	P-RRM	-10,864.17	21,901	4.1%
7	RRM	RAM	-11,440.91	23,054	3.6%
8	P-RRM	P-RRM	-10,720.51	21,613	30.3%
9	P-RRM	RAM	-11,082.22	22,337	22.3%
10	RAM	RAM	-12,139.71	24,452	0.0%
Best LL (LC models) - best LL (base model)			1,706.25		
Model averaging			-10,310.15		
Gain from model averaging			410.36		

the results for this dataset are in fact more in line with cases 10 and 11 from our simulated data analysis, implying that there is both taste and decision rule heterogeneity and more than two different models in the underlying DGP. Overall, these results are much in contrast with those of SP-1, as the gains obtained by model averaging implies evidence of decision rule heterogeneity.

5. Conclusions

In this paper, we revisit the use of latent class models to capture heterogeneity across decision-makers in behavioural processes such as attribute non-attendance and decision rule heterogeneity. These approaches have been very popular in recent years and have often been shown to produce significant gains in fit over simpler models. We argue that many such findings may be due to an unfair comparison with models not allowing for any heterogeneity and that the findings may in fact be driven by heterogeneity in the sensitivities to individual attributes rather than the presence of other phenomena. We have contrasted the findings obtained from such latent class models with those obtained using model averaging which combines the evidence from a number of separately estimated models. This latter approach of course leads to inferior model fit compared to a simultaneous latent class model as model averaging is based on combining different sample level models, i.e. using parameters that are appropriate at the sample level, but our findings provide some evidence that suggests that these bigger improvements may indeed be in part due to effects other than those that analysts seek to uncover. This is especially the case when showing that equivalent (or near equivalent) gains in model fit can be obtained from LC models that use the same structure in each class, thus only allowing for taste heterogeneity. In particular, there is little evidence of attribute non-attendance in either of our SP datasets. Whilst one dataset shows clear evidence of decision rule heterogeneity, the other does not.

In practice, an analyst should of course attempt to simultaneously allow for all different types of heterogeneity whilst remaining aware of potential confounding. This would however require the use of latent class structures with many different classes, which quickly become computationally and empirically infeasible. While we do not suggest that researchers abandon the use of latent class structures to investigate heterogeneity in behavioural processes, we urge for some caution in interpretation and suggest that model averaging can provide a useful tool for checking the likely validity of their insights. In particular, given that model averaging over similar models can result in a substantial improvement in model fit (as demonstrated by our second case study and by Hancock et al. 2020), a small improvement suggests that taste heterogeneity may be the driving factor behind a large gain (if observed) when moving to a latent class model.

As a closing comment, our results demonstrate that model averaging never does as well as fully flexible latent class models, even in the case showing clear decision rule heterogeneity. This suggests that there is more scope for heterogeneity in parameters across individuals conditional on a specific model structure rather than heterogeneity across individuals in the model structure itself. In many ways, this is not surprising given that datasets, especially from stated choice surveys, are relatively homogeneous in the structure of the choice sets and explanatory variables. This means that it is possibly unlikely that there would be substantial variation in how different individuals make choices in these scenarios, and consequently the models that explain these choices best are more likely to be dataset-specific rather than person-specific. More work is of course required, including testing using further simulated datasets as well as revealed preference datasets. This is especially important with a view to looking into the ability of model averaging to uncover heterogeneity of the type analysts increasingly attempt to uncover with latent class structures.

Acknowledgements

The authors would like to acknowledge the financial support by the European Research Council through the consolidator grant 615596-DECISIONS. The authors would also like to thank three anonymous reviewers for their comments on previous versions of this paper.

References

Abou-Zeid, M. and Ben-Akiva, M. (2010). A model of travel happiness and mode switching. In Hess, S. and Daly, A., editors, *Choice Modelling: The State-of-the-Art and the State-of-Practice*, pages 289–305. Emerald Publishing, UK.

Beck, M. J. and Hess, S. (2016). Willingness to accept longer commutes for better salaries: Understanding the differences within and between couples. *Transportation Research Part A: Policy and Practice*, 91:1–16.

Ben-Akiva, M. and Boccara, B. (1995). Discrete choice models with latent choice sets. *International Journal of Research in Marketing*, 12(1):9–24.

Boeri, M. and Longo, A. (2017). The importance of regret minimization in the choice for renewable energy programmes: Evidence from a discrete choice experiment. *Energy Economics*, 63:253–260.

Campbell, D., Lorimer, V., Aravena, C., and Hutchinson, W. G. (2010). Attribute processing in environmental choice analysis: implications for willingness to pay. 84th Annual Conference, March 29-31, 2010, Edinburgh, Scotland 91718, Agricultural Economics Society.

Charoniti, E., Kim, J., Rasouli, S., and Timmermans, H. J. (2020). Intrapersonal heterogeneity in car-sharing decision-making processes by activity-travel contexts: A context-dependent latent class random utility–random regret model. *International Journal of Sustainable Transportation*, pages

1–11.

Chorus, C. G. (2010). A new model of random regret minimization. *EJTIR, 10 (2), 2010.*

Chorus, C. G. (2014). Capturing alternative decision rules in travel choice models: a critical discussion. In *Handbook of choice modelling*. Edward Elgar Publishing.

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

Dey, B. K., Anowar, S., Eluru, N., and Hatzopoulou, M. (2018). Accommodating exogenous variable and decision rule heterogeneity in discrete choice models: Application to bicyclist route choice. *PloS one*, 13(11):e0208309.

Gopinath, D. (1995). *Modeling Heterogeneity in Discrete Choice Processes: Application to Travel Demand.* PhD thesis, MIT, Cambridge, MA.

Greene, W. H. and Hensher, D. A. (2003). A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research Part B: Methodological*, 37(8):681–698.

Hancock, T. O., Hess, S., Daly, A., and Fox, J. (2020). Using a sequential latent class approach for model averaging: Benefits in forecasting and behavioural insights. *Transportation Research Part A: Policy and Practice*, 139:429–454.

Hensher, D. (2014). Attribute processing as a behavioural strategy in choice making. In *Handbook of choice modelling*. Edward Elgar Publishing.

Hensher, D. A. (2010). Attribute processing, heuristics and preference construction in choice analysis. In Hess, S. and Daly, A. J., editors, *State-of Art and State-of Practice in Choice Modelling: Proceedings from the Inaugural International Choice Modelling Conference*, chapter 3, pages 35–70. Emerald, Bingley, UK.

Hensher, D. A. and Greene, W. H. (2010). Non-attendance and dual processing of common-metric attributes in choice analysis: a latent class specification. *Empirical Economics*, 39(4):413–426.

Hensher, D. A., Rose, J. M., and Greene, W. H. (2012). Inferring attribute non-attendance from stated choice data: implications for willingness to pay estimates and a warning for stated choice experiment design. *Transportation*, 39(2):235–245.

Hess, S. (2014). 14 latent class structures: taste heterogeneity and beyond. In *Handbook of choice modelling*, pages 311–329. Edward Elgar Publishing Cheltenham.

Hess, S., Beck, M., and Crastes dit Sourd, R. (2016). Can a better model specification avoid the need to move away from random utility maximisation? Transportation Research Board (TRB) 96th Annual Meeting.

Hess, S., Beck, M. J., and Chorus, C. G. (2014). Contrasts between utility maximisation and regret minimisation in the presence of opt out alternatives. *Transportation Research Part A: Policy and Practice*, 66:1–12.

Hess, S. and Rose, J. M. (2007). *A latent class approach to recognising respondents' information processing strategies in SP studies.* paper presented at the Oslo Workshop on Valuation Methods in Transport Planning, Oslo.

Hess, S., Shires, J., and Jopson, A. (2013a). Accommodating underlying pro-environmental attitudes in a rail travel context: Application of a latent variable latent class specification. *Transportation Research Part D*, 25:42–48.

Hess, S. and Stathopoulos, A. (2012). Linking the decision process to underlying attitudes and perceptions: a latent variable latent class construct. *paper presented at the* 13^{th} *International Conference on Travel Behaviour Research, Toronto.*

Hess, S. and Stathopoulos, A. (2013). A mixed random utility - random regret model linking the choice of decision rule to latent character traits. *Journal of Choice Modelling*, 9:27–38.

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

Transportation, 40(3):583–607.

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

Hole, A. R. (2011). A discrete choice model with endogenous attribute attendance. *Economics Letters*, 110(3):203–205.

Ishaq, R., Bekhor, S., and Shiftan, Y. (2013). A flexible model structure approach for discrete choice models. *Transportation*, 40(3):60–624.

Leong, W. and Hensher, D. A. (2014). Relative advantage maximisation as a model of context dependence for binary choice data. *Journal of choice modelling*, 11:30–42.

Montgomery, H. and Svenson, O. (1976). On decision rules and information processing strategies for choices among multiattribute alternatives. *Scandinavian Journal of Psychology*, 17(1):283–291.

Rezapour, M. and Ksaibati, K. (2021). Latent class model with heterogeneous decision rule for identification of factors to the choice of drivers' seat belt use. *Computation*, 9(4):44.

Scarpa, R., Gilbride, T., Campbell, D., and Hensher, D. A. (2009). Modelling attribute non-attendance in choice experiments for rural landscape valuation. *European Review of Agricultural Economics*, 36(2):151–174.

Smith, B., Goods, C., Barratt, T., and Veen, A. (2021). Consumer 'app-etite'for workers' rights in the australian 'gig'economy. *Journal of choice modelling*, 38:100254.

Swait, J. and Ben-Akiva, M. (1985). Incorporating random constraints in discrete models of choice set generation. *Transportation Research Part B*, 21(2):91–102.

Swärdh, J. and Algers, S. (2009). Willingness to accept commuting time for yourself and for your spouse: Empirical evidence from swedish stated preference data. *Working Papers - Swedish National Road & Transport Research Institute (VTI)*, 5.

Tversky, A. (1972). Elimination by aspects: a theory of choice. *Psychological Review*, 79:281–299.

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

Walker, J. and Ben-Akiva, M. (2002). Generalized random utility model. *Mathematical Social Sciences*, 43(3):303–343.