WHAT IS REALLY UNCOVERED BY MIXING DIFFERENT MODEL STRUCTURES: CONTRASTS BETWEEN LATENT CLASS AND MODEL AVERAGING

Thomas O. Hancock (Corresponding Author)
Choice Modelling Centre & Institute for Transport Studies
University of Leeds
tstoh@leeds.ac.uk

Stephane Hess
Choice Modelling Centre & Institute for Transport Studies
University of Leeds
S.Hess@its.leeds.ac.uk

ABSTRACT

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 them 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 find that this leads to significant reductions in the amount of heterogeneity of the type analysts have sought to uncover with latent class structures of late, with our results suggesting 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.

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

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) have also looked at

allowing for heterogeneity in the actual underlying model structure across individuals in a single sample. This has mainly made use of latent class structures, with two key applications, namely decision rule heterogeneity and in information processing work.

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 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. MNL, 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 MNL 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 MNL 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 MNL 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 MNL 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 MNL. 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 Random Utility Maximisation (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 MNL class 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 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 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. 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 for the choices observed for a set of N decision-makers is given by:

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

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). 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:

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

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 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:

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

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:

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

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

[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 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).

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 Mixed Multinomial Logit (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:

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

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.

3. SIMULATED DATA ANALYSIS

Before testing model averaging on our stated preference dataset, we use simulated datasets to look at the contrasting insights provided by different approaches when the true process is known. We first describe how we created our 10 different datasets before applying MNL, RRM, latent class models and model averaging.

3.1. Generation of simulated data

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}). For the MNL model, we define the utility as:

$$V_{nti} = \delta_i + \delta_{F_i} \cdot z_{F,n} + \beta_{TT} \cdot \alpha_{TT_i} \cdot TT_{nti} + \beta_{TC} \cdot \alpha_{IE,n} \cdot TC_{nti} + \beta_{AT} \cdot AT_{nti} + \varepsilon_{nti}, \tag{6}$$

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, $z_{F,n} = 1$ (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. For the RRM model, we define the regret for an alternative equivalently, replacing $\beta_x \cdot X_{nti}$ with $\sum_{j\neq i}^{4} log(1 + exp(\beta_x \cdot (X_{ntj} - X_{nti})))$. The constants which are added to the regret are also multiplied by -1 such that they have similar impacts on choice probabilities (as they do in MNL) by being applied in the same direction.

We use two sets of true parameter values for MNL, and two sets of true parameter values for RRM, with the parameter values given in Table 1. The first four datasets we create use the same data generation process (DGP) for all individuals and thus assume that all decision-makers use the same model with the same coefficients to make their choices (i.e. one dataset per column in Table 1). A further six datasets are created (one for each pair of sets of coefficients) using a random allocation such that half of the decision-makers use one DGP, and the other half use a different DGP. The types of heterogeneity in each dataset are given below in Table 2.

TABLE 1 : Coefficient values for the four 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 MNL1 and MNL2 such that they are very different. RRM1 and RRM2 have β-coefficients that are half the value of those of MNL1 and MNL 2 such that they have approximately the same scale.

Parameter	MNL1	RRM1	MNL2	RRM2
δ_C	-0.5000	-0.5000	1.0000	1.0000
δ_R	-1.5000	-1.5000	-0.5000	-0.5000
δ_A	-1.0000	-1.0000	1.0000	1.0000
δ_{F_C}	-0.5000	-0.5000	-0.2000	-0.2000
δ_{F_R}	0.5000	0.5000	1.5000	1.5000
δ_{F_A}	1.0000	1.0000	-0.5000	-0.5000
α_I	-0.5000	-0.5000	-0.3000	-0.3000
β_{TT}	-0.0040	-0.0020	-0.0050	-0.0025
β_{TC}	-0.0280	-0.0140	-0.0100	-0.0050
β_{AT}	-0.0080	-0.0040	-0.0120	-0.0060
α_{TT_R}	1.2500	1.2500	0.8000	0.8000
α_{TT_A}	2.0000	2.0000	1.7000	1.7000
α_{TT_H}	1.5000	1.5000	1.7000	1.7000
Share(Car)	53.06%	54.80%	43.12%	48.70%
Share(Rail)	9.54%	7.70%	6.36%	5.32%
Share(Air)	13.00%	11.76%	39.60%	33.66%
Share(HSR)	24.40%	25.74%	10.92%	12.32%

3.2. Results from simulated datasets

For each of the simulated datasets, we estimate five different models. The first two of these are basic MNL and basic RRM models. The next three are latent class models with two classes, using

TABLE 2 : Types of heterogeneity in each dataset

Dataset	DGP 50%	DGP 50%	Heterogeneity type Taste	Heterogeneity type Decision rule
1	MNL1		no	no
2	MNL2		no	no
3	RRM1		no	no
4	RRM2		no	no
5	MNL1	MNL2	yes	no
6	MNL1	RRM1	no	yes
7	MNL1	RRM2	yes	yes
8	MNL2	RRM1	yes	yes
9	MNL2	RRM2	no	yes
10	RRM1	RRM2	yes	no

all three combinations of models, i.e. MNL-MNL, MNL-RRM and RRM-RRM. This means that for each dataset, we have two models that test for taste heterogeneity alone (MNL-MNL and RRM-RRM) and one model that allows for taste heterogeneity and decision rule heterogeneity (MNL-RRM). As we have all variations of datasets that include taste and/or decision rule heterogeneity covered by our 10 datasets, 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 each time highlight the best BIC for each DGP, along with the noting whether there is a clear “winner" in model averaging or a more even split between models.

We first look at the four datasets which come from a single model without heterogeneity (i.e. models 1-4). We note that each time, the model using the correct decision rule outperforms its counterpart, while none of the three LC structures can reject the single class model. The differences in fit between MNL and RRM are small, in line with many previous findings in the literature, at least with the most common RRM implementation.

We notice that there are four datasets for which there is clear evidence of heterogeneity, with datasets 5, 7, 8 and 10 having improvements in model fit of at least 150 log-likelihood units by moving from a basic MNL or basic RRM model to a latent class model. These are the four datasets where we either mix two models of the same type (5 and 10) or where the models of different types also use substantially different relative taste coefficients (7 and 8, noting that the parameters for MNL1 and RRM1 are consistent with each other, as are MNL2 and RRM2). Crucially, we also see that for each of the four cases, the LC structure that is in line with the DGP performs best. However, the differences in fit are very small, again reiterating the point about similarities between MNL and RRM.

We finally look at the two models that combine different decision rules but use consistent parameters in the two, i.e. focussing on decision rule rather than taste heterogeneity. These are

TABLE 3 : Results from MNL, RRM and latent class models for each of the 10 simulated datasets, together with the results from model averaging.

Dataset	1	2	3	4	5	6	7	8	9	10
Data creation:	MNL1	MNL2	RRM1	RRM2	MNL1 MNL2	MNL1 RRM1	MNL1 RRM2	MNL2 RRM1	MNL2 RRM2	RRM1 RRM2
Log-likelihoods (LL(0) = -6,931.47)										
MNL	-4,816.03	-5,043.07	-4,641.41	-4,986.07	-5,357.92	-4,736.56	-5,227.40	-5,293.52	-5,037.90	-5,203.48
RRM	-4,841.62	-5,046.22	-4,631.31	-4,981.98	-5,368.51	-4,743.00	-5,239.64	-5,299.49	-5,038.91	-5,203.23
MNL-MNL	-4,804.49	-5,028.90	-4,637.99	-4,976.56	-5,101.61	-4,728.18	-5,047.94	-5,040.73	-5,021.55	-5,060.08
MNL-RRM	-4,807.08	-5,039.74	-4,622.68	-4,976.86	-5,114.58	-4,727.61	-5,046.22	-5,036.22	-5,027.76	-5,055.41
RRM-RRM	-4,828.56	-5,034.32	-4,625.27	-4,969.88	-5,115.40	-4,734.96	-5,060.76	-5,041.17	-5,027.61	-5,052.90
LC2-LC1	11.53	14.17	8.63	12.10	256.31	8.95	181.17	257.30	16.35	150.32
Bayesian Information Criterion (BIC)										
MNL	**9,742.77**	**10,196.86**	9,393.54	10,082.87	10,826.56	**9,583.84**	10,565.51	10,697.76	**10,186.52**	10,517.68
RRM	9,793.96	10,203.17	**9,373.34**	**10,074.68**	10,847.75	9,596.72	10,589.99	10,709.70	10,188.55	10,517.18
MNL-MNL	9,838.95	10,287.77	9,505.95	10,183.09	**10,433.17**	9,686.33	10,325.84	10,311.42	10,273.06	10,350.12
MNL-RRM	9,844.12	10,309.45	9,475.32	10,183.68	10,459.12	9,685.19	**10,322.41**	**10,302.41**	10,285.49	10,340.79
RRM-RRM	9,887.07	10,298.61	9,480.50	10,169.71	10,460.76	9,699.89	10,351.48	10,312.30	10,285.18	**10,335.77**
Model Averaging Share										
MNL	**99.93%**	**87.03%**	13.97%	9.57%	0.00%	**71.96%**	0.04%	0.00%	**60.92%**	0.01%
RRM	0.07%	12.97%	**86.03%**	**90.43%**	0.00%	**28.04%**	0.00%	0.00%	**39.08%**	0.00%
MNL-MNL	-	-	-	-	**99.98%**	-	9.55%	15.18%	-	10.29%
MNL-RRM	-	-	-	-	0.01%	-	**90.34%**	**84.52%**	-	2.04%
RRM-RRM	-	-	-	-	0.01%	-	0.07%	0.30%	-	**87.67%**
Model Averaging Log-likelihood										
LL	-4,816.03	-5043.00	-4631.02	-4981.93	-5,101.61	-4,735.38	-5,046.20	-5,036.08	-5,037.17	-5,052.79
Improvement	0.00	0.07	0.29	0.04	0.00	1.17	0.02	0.14	0.73	0.11

models 6 and 9. For these two cases, the improvements in model fit for the latent class structures are very modest, to the point of not passing a statistical test for rejecting the single class model. In fact, the improvements found for datasets 6 and 9 are comparable to those found by datasets 1-4. This finding again suggests that MNL and RRM are very similar models when using parameters that are consistent across structures. A small gain in log-likelihood by moving to latent class models results in a worse BIC value, and thus we observe that the best BIC for datasets without taste heterogeneity is found by either the basic MNL model (for datasets generated by a single MNL or by MNL and RRM) or the basic RRM model (for datasets generated by a single RRM). We can thus reject the latent class models in these cases and apply model averaging over only the basic MNL and basic RRM models. For models with taste heterogeneity, however, we apply model averaging across all latent class models as well as the MNL and RRM models.

The results from model averaging are also given in Table 3. For the first four datasets, we observe a large share (minimum 86%) given to the model that created the data, as we would expect. For the two remaining datasets without taste heterogeneity (datasets 6 and 9), we observe a more equal split in the shares, thus implying the presence of decision rule heterogeneity. The results for the datasets with taste heterogeneity also clearly match the data generation process. MNL-MNL obtains almost the full share for the dataset created by two MNLs, and RRM-RRM obtains 88% of the share for the dataset created by two RRMs. Additionally, we observe large shares for the MNL-RRM latent class model (90% and 85%, respectively) for the two datasets created by MNL and RRM models where there was also taste heterogeneity. It is crucial here to note that applying model averaging across all latent class models in all cases may result in unclear results, as latent class models without substantial improvements in model fit may only be picking up noise in the data. Thus averaging across these models will highlight models that only explain noise, rather than the underlying data generation process. Whilst we may see a greater improvement in model fit through applying model averaging across larger sets of candidate models utilising a wider range of decision rules, the results here demonstrate that model averaging can help differentiate some cases where taste heterogeneity alone exists. Additionally, model averaging helps us when taste and decision rule heterogeneity exist, with only small improvements in model fit for the MNL-RRM model (1.72 and 4.51 log-likelihood units in DGP 7 and 8, respectively) over the MNL-MNL model, but clear advantages for the MNL-RRM model in terms of the model averaging shares.

4. ANALYSIS ON SP DATA

This section presents our work on a typical SP dataset. 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 a SC dataset from Hess and Stathopoulos (2013) where 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.

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).

4.2.1. Specification

We start by estimating a simple MNL model, where we use a logarithmic transform on the fare attribute given earlier evidence of strong non-linearity. This 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).

We next move to the latent class model for attribute non-attendance. We use a model 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

$$\pi_s = \prod_{k=1}^{K} \left(\Lambda_{s,k} \left(1 - P_{\text{NA},k}\right) + \left(1 - \Lambda_{s,k}\right) P_{\text{NA},k} \right), \tag{7}$$

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 $P_{\text{NA},k}$ is the probability of non-attendance to attribute k, thus $P_{\text{NA},k} = \frac{e^{\delta_k}}{e^{\delta_k}+1}$), 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 individual models, corresponding to all possible combinations of attribute attendance and non-attendance, i.e. 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 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.

4.2.2. Results

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

TABLE 4 : MNL results for public transport route choice

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
β_{tt}	-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 confirmatory LC 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 confirmatory latent class model for ANA. We also see that the marginal utility parameters, which now only apply in the attendance classes, have increased substantially compared to the base model. This is in line with the notion that the MNL model would find an intermediary value between 0 for the non-attenders and a positive value for those attending to the attribute. However, the implied rates of non-attendance are unrealistically high, exceeding 50% for all attributes except fare.

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 MNL 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

TABLE 5 : Confirmatory latent class model for attribute non-attendance

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	0.329	4.23
β_{tt}	-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
$\delta_{\text{NA,tt}}$	0.3232	1.11
$\delta_{\text{NA,log-fare}}$	-0.5142	-3.43
$\delta_{\text{NA,crowding}}$	0.7767	3.3
$\delta_{\text{NA,rate of delays}}$	0.7363	2.43
$\delta_{\text{NA,average delay}}$	1.1917	4.02
$\delta_{\text{NA,info system charged}}$	3.1776	3.82
$\delta_{\text{NA,info system free}}$	0.9874	3.61
	Implied rate of NA	
	Estimate	Rob.t.ratio(0)
travel time	0.5801	8.18
fare	0.3742	10.65
crowding	0.685	13.49
rate of delays	0.6762	10.21
average delay	0.767	14.48
info system charged	0.96	30.05
info system free	0.7286	13.47

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 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,

TABLE 6 : Model averaging for ANA work

LL(final) = -3,363.28, LL(0) = -4,042.89			Implied rate of NA		
	Estimate	Rob.t.ratio(0)		Estimate	Rob.t.ratio(0)
$\delta_{\text{NA,average delay}}$	-1.9099	-1.95	average delay	0.129	1.17
$\delta_{\text{NA,ch inf sys}}$	0.0844	0.05	info system charged	0.5211	1.22
$\delta_{\text{NA,free inf sys}}$	-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)
β_{tt}	-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)	-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
$\beta_{\text{ch inf sys}}$	0	-0.08 (-1.04)	0	-0.27 (-3.67)	0	-0.04 (-0.57)	0	-0.24 (-3.24)
$\beta_{\text{free inf sys}}$	0.36 (5.96)	0.34 (5.06)	0	0	0.36 (5.91)	0.35 (5.22)	0	0

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.

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.

4.3.1. Specification

To maximise the possibility of finding such heterogeneity, we specifically choose to consider five very different decision rules, namely:

Multinomial logit (MNL): We assume that the utility a respondent n obtains from alternative i (out of J alternatives) in choice task t is:

$$V_{nti} = U_{nti} + \varepsilon_{nti}, \tag{8}$$

where V_{nti} and ε_{nti} are the deterministic and random components of utility respectively. The assumption of a type I extreme value distribution for ε_{nti} then gives us the usual MNL choice probabilities:

$$P_{MNL,nti} = \frac{e^{V_{nti}}}{\sum_{j=1}^{J} e^{V_{ntj}}}. \tag{9}$$

Random regret minimisation (RRM): We base our random regret minimisation (RRM) model on the updated specification of Chorus (2010). Thus, the deterministic regret for respondent n for alternative i in choice task t is given by:

$$R_{nti} = \delta_{RRM,i} + \sum_{j \neq i}^{J} \sum_{k=1}^{K} ln(1 + e^{\beta_k (x_{ntjk} - x_{ntik})}) \tag{10}$$

where $k = 1, ..., K$ is an index across attributes, β_k is an attribute-specific coefficient for attribute k and $\delta_{RRM,i}$ is an alternative specific constant. With the error component of regret also being given by a type I extreme value distribution, the corresponding RRM probabilities for a respondent n choosing alternative i in choice task t is:

$$P_{RRM,nti} = \frac{e^{-R_{nti}}}{\sum_{j=1}^{J} e^{-R_{ntj}}} \tag{11}$$

Decision field theory (DFT): DFT is a dynamic, stochastic model where the preferences for alternatives update over the course of the decision-making process (Busemeyer and Townsend, 1992). Whilst it has not traditionally been used to model travel behaviour, work by Roe et al. (2001) extended the model such that it can be applied to multi-attribute, multi-alternative choice tasks and Berkowitsch et al. (2014) demonstrated that DFT can outperform logit and probit models for consumer choice tasks. More recently, Hancock et al. (2018) and Hancock et al. (2020c) have demonstrated that DFT can outperform standard models for typical travel behaviour data. It is mathematically a very different model to MNL or RRM as it assumes that a decision-maker stochastically considers the different attributes of the alternatives over the course of a decision-making process for a single choice task. In each of these steps, a single attribute is considered - note that this is different from ANA in that all attributes are attended, but one at a time, and with different attributes being attended to different numbers of times as a function of their importance. This results in the preference values updating iteratively:

$$P_\tau = S \cdot P_{\tau-1} + V_\tau, \tag{12}$$

where P_τ is a column vector containing the preference values of each alternative i at time τ. S is a feedback matrix with memory and sensitivity parameters[2] and V_τ is a valence vector, which is the preference accumulated in the step τ and will depend on which attribute is attended to in that step. The valence vector can be described by:

$$V_\tau = C \cdot M \cdot W_\tau + \varepsilon_t \tag{13}$$

where C is a contrast matrix used to rescale the values such that they total zero, M is the matrix of attribute values and $W_\tau = [0..1..0]'$ with entry $k = 1$ if and only if attribute k is the attribute being attended to by the decision-maker at preference updating step τ. A DFT model thus estimates a weight, w_k, for the likelihood of attending to attribute k. As the error term, ε_τ is drawn from a normal distribution with mean 0 (and a variance which is an estimated parameter), the preference values P_τ converge to a multivariate normal distribution. To calculate the probabilities of alternatives under decision field theory we thus simply require the expectation and covariance of P_τ (ξ_τ and Ω_τ, respectively). Hence the probability of choosing alternative i from a set of J alternatives at time τ is (dropping the index for choice set and individual):

$$P_{DFT,i}\left[\max_{j \in J} P_\tau[j] = P_\tau[i]\right] = \int_{X>0} exp\left[-(X-\Gamma)'\Lambda^{-1}(X-\Gamma)/2\right] / (2\pi|\Lambda|^{0.5}) dX \tag{14}$$

[2] We implement the feedback matrix equivalently to Hancock et al. (2020c).

with $X = [P_\tau[j] - P_\tau[1], ..., P_\tau[j] - P_\tau[J]]'$, $\Gamma = L\xi_\tau$, $\Lambda = L\Omega_\tau L'$ and L a matrix comprised of a column vector of 1s and a negative identity matrix of size $J-1$ where J is the number of alternatives. The column vector of 1s is placed in the i^{th} column where i is the chosen alternative. The DFT model utilised in the empirical tests in this paper is based on the version in Hancock et al. (2020c), which incorporates attribute-specific scaling factors.[3]

Quantum amplitude model (QAM) Our quantum model is based on the quantum amplitude model defined by Hancock et al. (2020a). Under a quantum model, the possible choice alternatives can be represented by a set of orthogonal vectors which make up the basis for a multidimensional Hilbert space (Bruza et al., 2015). A decision-maker's opinion or 'belief state' can then be represented by another vector within this space. The action of making a choice is represented by a projection from the belief state vector onto the vector representing the chosen alternative (see figures in Hancock et al. 2020a). Setting the belief state vector to be of unit length results in the set of squared 'lengths' (amplitudes) of these projection onto each of the vectors representing the possible alternatives summing to one. This means that, rather than the probability of alternatives being based on utilities with some error terms, they are based on the belief state vector. Each alternative has an amplitude rather than a utility, and the error is inherent in the belief state vector itself, which directly captures the fact that a decision-maker may choose any of the alternatives. Evidence in favour of an alternative increases the relative 'lengths' of the vector for the alternative, resulting in the probability for this alternative increasing. Consequently, the belief state vector captures the fact that a decision-maker may choose different alternatives without the need for an additional error term. There are many reasons for adopting quantum frameworks Bruza et al. 2015), but most importantly in the context of travel behaviour analysis, they are flexible models whereby different value functions can be integrated into the structure together with 'quantum rotations' or 'changes in perspectives' for a change in choice context. In particular, Hancock et al. (2020a) demonstrate that the quantum amplitude model can outperform traditional choice models for standard travel behaviour data. We thus include it as an additional model in this paper to further broaden the possibility of finding decision rule heterogeneity, given that it is, like DFT, a mathematically very different choice model to MNL or RRM.

Under QAM, there are a number of possible functions that can be used to define the amplitudes for each alternative, with Hancock et al. (2020a) giving a detailed overview and empirical tests of these possibilities. In this case, we consider amplitudes based on 'softplus' (random regret-like) functions. For alternative i (for respondent n in choice task t), we define the amplitude:

$$\psi_{nti} = \left(\delta_{QAM,i} + I_0 + \sum_{j \neq i}^{J} \sum_{k=1}^{K} wt_{ij} \cdot ln(1 + e^{\beta_k (x_{ntik} - x_{ntjk})}) \right) / \sqrt{\mathscr{N}}, \tag{15}$$

where $\delta_{QAM,i}$ are alternative-specific constants, I_0 is a constant that has the same value across all alternatives, wt_{ij} is a weight for the relative importance of the comparison between alternatives i and j and β_k is a coefficient for attribute k as before for RRM. $\mathscr{N}$ is a normalisation

[3] For a full description of decision field theory, how it can be applied and how the different parameters in the model work, readers should consult Hancock et al. (2018) and Hancock et al. (2020c).

factor, which ensures that the probabilities with which each alternative is chosen sum to one. It is obtained from the sum of the squared moduli:

$$\mathscr{N} = \sum_{i}^{J} \left| \left(\delta_{QAM,i} + \sum_{j \neq i}^{J} \sum_{k=1}^{K} wt_{ij} \cdot ln(1 + e^{\beta_k (x_{ntik} - x_{ntjk})}) \right) \right|^2, \tag{16}$$

where $i = 1, ...J$ is an index across the possible alternatives. Once these amplitudes have been calculated, the application of the normalisation means that the probability for each alternative[4] can be defined simply as:

$$P_{QAM,jnt} = \psi_{jnt}^2. \tag{17}$$

Relative advantage maximisation (RAM) In RAM (Leong and Hensher, 2014), the utility for respondent n in choice task t is:

$$U_{nti} = \delta_{RAM,i} + \sum_{k=1}^{K} \beta_k \cdot x_{ntik} + \sum_{j \neq i} RA(i,j), \tag{18}$$

which is equivalent to a multinomial logit model with the addition of the comparison of relative advantages $RA(i,j)$ of alternative i with each of the other alternatives. This relative advantage is then defined as:

$$RA(i,j) = \frac{A(i,j)}{A(i,j) + D(i,j)}, \tag{19}$$

where the advantages are calculated $A(i,j) = ln(1 + e^{\beta_k (x_{ntik} - x_{ntjk})})$ and the disadvantages $D(i,j) = ln(1 + e^{\beta_k (x_{ntjk} - x_{ntik})})$.

4.3.2. Results

For our SP dataset, we first apply the five different models individually, obtaining the results given in Table 7. We see that DFT obtains the best log-likelihood ahead of QAM, while the performance of the three logit-style models is noticeably poorer and comparatively more similar. As a first step, we look at model averaging across these five individual models with different decision rules, where the resulting shares and fit are shown in Table 7. We see that the model average leads to a further small improvement in model fit over the best fitting individual model, i.e. DFT, where this model also obtains by far the largest share in the model average. As with earlier examples, the shares are not necessarily proportional to the model fit of the individual model, and we see that RRM obtains a substantially larger share than QAM, despite having poorer overall individual log-likelihood. This again shows that some models can work well for some people even if they obtain a lower overall fit to the sample.

In practice, the estimation of a latent class model with five separate classes all using individual decision rules is computationally challenging and most applications rely on just combining a

[4]Note that as we are using real values only in the specification of the amplitudes, squaring these amplitudes for the probability suffices. If we had complex valued amplitudes, we would require the use of complex conjugates.

TABLE 7 : Results from different individual models applied to the SP dataset

Model	Type	Log-likelihood	BIC	MA Share
1	MNL	-3,360.43	6,803	0.00%
2	RRM	-3,363.91	6,810	17.67%
3	DFT	-3,317.18	6,749	76.54%
4	QAM	-3,336.44	6,771	5.70%
5	RAM	-3,354.55	6,791	0.08%
Model averaging		-3,312.40		

couple of different rules. We therefore look at the estimation of 15 different latent class structures with two classes each, picking all combinations of two model structures with replacement, thus also allowing for five models where the two classes are of the same type, i.e. looking for taste heterogeneity alone. Table 8 gives the log-likelihoods of these models. For all 15 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. 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 8 : Results from latent class models applied to SP dataset

Model	Class 1	Class 2	Log-likelihood	BIC	MA1 Share	MA2 Share
1	MNL	MNL	-3,113.13	6,399	0.0%	0.0%
2	MNL	RRM	-3,102.66	6,378	0.0%	-
3	MNL	DFT	-3,099.84	6,380	6.7%	-
4	MNL	QAM	-3,106.76	6,394	0.0%	-
5	MNL	RAM	-3,100.79	6,374	0.0%	-
6	RRM	RRM	-3,106.33	6,385	16.0%	22.8%
7	RRM	DFT	-3,086.79	6,354	11.7%	-
8	RRM	QAM	-3,096.35	6,373	0.0%	-
9	RRM	RAM	-3,104.22	6,381	0.0%	-
10	DFT	DFT	-3,077.79	6,361	52.8%	62.3%
11	DFT	QAM	-3,085.28	6,376	0.0%	-
12	DFT	RAM	-3,085.38	6,351	0.0%	-
13	QAM	QAM	-3,095.71	6,380	12.8%	14.9%
14	QAM	RAM	-3,094.59	6,370	0.0%	-
15	RAM	RAM	-3,100.27	6,373	0.0%	0.0%
MA 1 (all 15)			-3,071.46			
MA 2 (1,6,10,13,15)			-3,071.65			

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 DFT, which is the best performing individual model in Table 7, combining the model with a different structure does not reach as high a log-likelihood as a structure with two separate DFT classes, although a better BIC may be obtained. 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 is outperforming the two LC models that use the same model type in both classes. The MNL-RRM model outperforms RRM-RRM by 3.67 log-likelihood units and outperforms MNL-MNL by 10.47 units. Additionally, QAM-RAM has a better log-likelihood than either QAM-QAM or RAM-RAM.

Further evidence is given in the model averaging results in Table 8. Overall, we observe that averaging across all 15 latent class models results in a slightly better log-likelihood than the best performing individual model (DFT-DFT). Single model type classes also obtain 81.6% of the weight in model averaging. We also observe that averaging across the 5 models that only capture taste heterogeneity alone obtains a very similar log-likelihood in model averaging. These findings again highlight the importance of within-model taste heterogeneity, at least for this data.

We explore the most common example of decision rule heterogeneity (MNL-RRM) in more detail by also considering the outputs for the parameter estimates, in comparison to a model average performed on MNL and RRM. The results for this are shown in Table 9. For each model, we have coefficients for travel time (TT), the log of fare (LFare[5]), rate of crowding (Crowd), length of delays (Delay), rate of delays (Rate), a reliability level (Rel), created by calculating the expected length of delays, and the provision of a charged delay information service (Inf) or a free service (InfF). Finally, we include two alternative specific constants for the first two alternatives.

Table 9 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 MNL 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 MNL model on its own, with an improvement of just 0.07 log-likelihood units. As a contrast, the latent class approach results in a vast improvement in model fit (258 units). At face value, this would again suggest decision rule heterogeneity, although the fit is not much better than for the MNL-MNL 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 MNL-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

[5]Note that we use a logarithmic transform of the fare rather than the fare itself as a cost damping effect is observed.

TABLE 9 : A detailed example of model averaging compared to a simultaneous latent class approach using MNL and RRM

	Latent Class - 1 model 21 pars, estimated simultaneously		Model averaging - 3 models 2*10 pars, then 1 for MA	
	Class 1:MNL	Class 2:RRM	Class 1: MNL	Class 2: RRM
Class LL: Log-likelihood	-3,645.30	-4,431.55	-3,360.43	-3,363.91
	-3,102.66		-3,360.36	
δ_{alt1}	0.64 (6.42)	0.04 (0.27)	0.39 (5.85)	0.27 (4.17)
δ_{alt2}	0.25 (2.81)	0.20 (1.13)	0.16 (3.3)	0.17 (3.38)
β_{TT}	-0.05 (-6.74)	-0.05 (-6.79)	-0.05 (-9.5)	-0.03 (-9.58)
β_{Lfare}	-3.21 (-6.1)	-11.32 (-7.58)	-6.00 (-18.87)	-4.11 (-17.66)
β_{Crowd}	-0.31 (-7.41)	-0.15 (-2.89)	-0.22 (-8.58)	-0.15 (-8.59)
β_{Delay}	-0.06 (-1.27)	-0.05 (-1.29)	-0.03 (-3.24)	-0.02 (-3.06)
β_{Rate}	-0.34 (-4.82)	-0.09 (-1.76)	-0.19 (-5.96)	-0.12 (-5.82)
β_{Rel}	-0.05 (-3.22)	0.00 (0.06)	-0.06 (-2.64)	-0.04 (-2.71)
β_{Inf}	-0.10 (-0.82)	-0.16 (-1.09)	-0.09 (-1.13)	-0.05 (-0.95)
β_{InfF}	0.54 (5.84)	0.05 (0.47)	0.33 (4.95)	0.22 (4.85)
π_m	59.30% (10.89)	40.70%	87.70% (2.7)	12.30%

far superior model fit. Together with the poor improvement from model averaging, these results suggest that most of the model improvement is due to taste rather than 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 first 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 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 and 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 this context, model averaging can provide useful tests. Small improvements in model fit from model averaging point towards other sources of heterogeneity, as our previous work (Hancock et al., 2020b) demonstrates that large improvements from model averaging can be found when averaging over very similar models.

As a closing comment, the findings in the application looking at decision rule heterogeneity are particularly insightful. They suggest 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. The models that work best are more likely to be dataset-specific rather than person-specific. More work is of course required, including testing using simulated 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.

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.

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

Berkowitsch, N. A., Scheibehenne, B., and Rieskamp, J. (2014). Rigorously testing multialternative decision field theory against random utility models. *Journal of Experimental Psychology: General*, 143(3):1331.

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.

Bruza, P. D., Wang, Z., and Busemeyer, J. R. (2015). Quantum cognition: a new theoretical approach to psychology. *Trends in cognitive sciences*, 19(7):383–393.

Busemeyer, J. R. and Townsend, J. T. (1992). Fundamental derivations from decision field theory. *Mathematical Social Sciences*, 23(3):255–282.

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.

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., Broekaert, J., Hess, S., and Choudhury, C. F. (2020a). Quantum probability: a new method for modelling travel behaviour. *Submitted*.

Hancock, T. O., Hess, S., and Choudhury, C. F. (2018). Decision field theory: Improvements to current methodology and comparisons with standard choice modelling techniques. *Transportation Research Part B: Methodological*, 107:18–40.

Hancock, T. O., Hess, S., Daly, A., and Fox, J. (2020b). Using a sequential latent class approach for model averaging: benefits in forecasting and behavioural insights. *Submitted*.

Hancock, T. O., Hess, S., Marley, A. A. J., and Choudhury, C. F. (2020c). An accumulation of preference: two alternative dynamic models for understanding transport choices. *Submitted*.

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. 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.

Roe, R. M., Busemeyer, J. R., and Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review*, 108(2):370.

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.

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.

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

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