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, which have traditionally been used for taste heterogeneity, are increasingly used as a tool for capturing heterogeneity in other components, 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. These claims have not been without criticism highlighting the potential risk of confounding with other more model-specific heterogeneity. In this paper, we add a different angle of thought to this conversation 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.

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

1. INTRODUCTION

Latent class structures have long been used as a tool for introducing heterogeneity across individual decision makers in choice models (Greene and Hensher, 2003; Hess, 2014). Over the last decade, there has also been increasing interest in using the models to allow for heterogeneity in the actual underlying model structure across individuals, with two key applications, in decision rule heterogeneity and in information processing work. 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. Hole, 2011; Scarpa et al., 2009; Hensher and Greene, 2010; Campbell et al., 2010; 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 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 likelihood for decision maker n is then given by:

$$L_n(\beta, \pi) = \sum_{s=1}^{S} \pi_s \prod_{t=1}^{T} P_{ni^*t}(\beta_s = \beta \circ \Lambda_s). \tag{1}$$

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 $L_n(\beta_m, m)$ give the probability of the observed sequence of choices for decision maker n, conditional on using a choice model identified as m, where this uses a vector of parameters β_m. The Hess et al. (2012) framework is based on the idea that M different behavioural processes are used in the data. The probability for the sequence of choices observed for decision maker n is now given by:

$$L_n(\beta, \pi) = \sum_{m=1}^{M} \pi_{n,m} L_n(\beta_m, m), \tag{2}$$

where we use different behavioural processes in different classes, with the probability of decision rule class m for decision maker n given by $\pi_{n,m}$. Hess et al. (2012) additionally allow for random heterogeneity in parameters within individual decision rule classes, such that:

$$L_n(\Omega, \pi) = \sum_{m=1}^{M} \pi_{n,m} \int_{\beta_m} L_n(\beta_m, m) f(\beta_m, \Omega_m) \, \mathrm{d}\beta_m, \tag{3}$$

where $\beta_m \sim f(\beta_m, \Omega_m)$ and $\Omega_m = \langle \Omega_1, \ldots, \Omega_M \rangle$.

Hess et al. (2012) use the model to allow for mixtures between random utility maximisation, random regret minimisation and elimination by aspects. 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.

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).

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 in fit and apparent meaningful insights (see references above), it has also not been without criticism. In particular, concerns have been raised about extensive risk of confounding between taste heterogeneity 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. There is then the real possibility that evidence of a substantial class allocation probability for different classes will be driven by heterogeneity in sensitivities rather than actual 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, as in the work of Hess et al. (2013b) and Hess et al. (2016) is computationally very difficult. In the present paper, we thus use a different approach by highlighting how model averaging can be used as a diagnostic tool for the potential confounding between taste heterogeneity and other heterogeneity.

Model averaging, in this context, can be implemented as a sequential latent class model. Whereas a fully flexible model simultaneously estimates the parameters of the class component models as well as the class shares, a model averaging approach estimates the separate classes as individual models first, before estimating the class shares separately with the individual model parameters fixed. To apply model averaging, we first estimate a number of different individual models, where say $L(C_n \mid m, \Omega_m)$ gives the likelihood of the sequence of choices C_n observed for person n, conditional on using model m, where this model uses a vector of parameters Ω_m. We have that:

$$L(C_n \mid m, \Omega_m) = \int_{\beta_m} \prod_{t=1}^{T_n} P_m\left(j^*_{n,t} \mid \beta_m\right) f_m\left(\beta_m \mid \Omega_m\right) \mathrm{d}\beta_m. \tag{4}$$

In this general notation, we have that $P_m\left(j^*_{n,t} \mid \beta_m\right)$ gives the probability of the observed choice $j^*_{n,t}$ for decision maker n in choice situation t, conditional on using model m, where we allow for a general notation such that the parameters β_m are distributed according to $f_m\left(\beta_m \mid \Omega_m\right)$. Of course, it is possible that no random heterogeneity is used in which case the integral drops out, or that a latent class structure is used, replacing the integral with a weighted summation.

An analyst will estimate M different such individual models, of differing form, each yielding

a set of parameters and a likelihood at the individual level $L(C_n \mid m, \Omega_m)$. In the context of the present paper, the set of M models would include models with different specifications for IPS or different specifications in terms of underlying decision rule. The model averaging process then computes the overall likelihood for person n as the weighted average across M models, with:

$$L_n(\pi_n, \Omega) = \sum_{m=1}^{M} \pi_{m,n} L(C_n \mid m, \Omega_m), \tag{5}$$

where $\sum_{m=1}^{M} \pi_{m,n} = 1$ and $0 \leq \pi_{m,n} \leq = 1$. This overall likelihood is conditional on the vector of weights $\pi_n = \langle \pi_{1,n}, \ldots, \pi_{M,n} \rangle$ and the combined parameter estimates from the different models $\Omega = \langle \Omega_1, \ldots, \Omega_M \rangle$.

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.

The remainder of this paper is organised as follows. We first present the data used in our analysis (Section 2). This is followed in Section 3 by our work on attribute noon attendance, and Section 4 by our work on decision rule heterogeneity. Finally, some conclusions are presented in Section 5.

2. DATA

Our main analysis relies on a SC dataset 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. 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 cost and availability of a delay information service (in £). 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.

3. INFORMATION PROCESSING WORK

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

We first estimate 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). The results for this model are shown in Table 1 where all estimates are of the correct sign.

TABLE 1 : MNL results for public transport route choice

LL(final)	-3,366.95	
ρ^2	0.1672	
adj. ρ^2	0.165	
	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

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 use of a simplified approach, by instead setting

$$\pi_s = \prod_{k=1}^{K} \left(\Lambda_{s,k} \left(1 - P_{\text{N-A},k}\right) + \left(1 - \Lambda_{s,k}\right) P_{\text{N-A},k} \right), \tag{6}$$

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 estimated K separate δ elements (with $P_{\text{N-A},k} = \frac{e^{\delta_k}}{e^{\delta_k}+1}$), as opposed to $2^K - 1$, leading to significant reductions in the number of parameters.

The results for this model are shown in Table 2. 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 parameters in the attendance classes have increased substantially, where 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.

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 to one with the two ASCs only.

TABLE 2 : Confirmatory latent class model for attribute non-attendance

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

We then estimate the model averaging structure, where we again use multiplicative class allocation probabilities, as in the LC model. We initially estimate seven class allocation weights as in the LC model but find that four 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 3. We see that this model now only offers a marginally better log-likelihood than the MNL model in Table 1, much in contrast with the LC model in Table 2. 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 see low rates for average delay and the free information system, with a higher rate for the charged system.

TABLE 3 : Model averaging for ANA work

LL(final)	-3,363.28		Implied rate of NA		
	Estimate	Rob.t.ratio(0)		Estimate	Rob.t.ratio(0)
$\delta_{\text{NA,average delay}}$	1.1917	4.02	average delay	0.129	1.17
$\delta_{\text{NA,ch inf sys}}$	3.1776	3.82	info system charged	0.5211	1.22
$\delta_{\text{NA,free inf sys}}$	0.9874	3.61	info system free	0.2399	2.33

	Information for 8 retained models							
individual 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	2	1	6	3	5	4	8	7
best fitting N	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 (-8.5)	-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

We further see that the 8 models that obtain the best individual log-likelihoods are also the only 8 models that contribute to the model average. However, the best ranking model individually is not necessarily the one contributing the most to the model average. Finally, 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 homogenous 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.

4. DECISION RULE HETEROGENEITY WORK

We next turn to decision rule heterogeneity. To maximise the possibility of finding such heterogeneity, we 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:

$$V_{int} = U_{int} + \varepsilon_{int}, \tag{7}$$

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

$$P_{MNL,int} = \frac{e^{V_{int}}}{\sum_{j=1}^{J} e^{V_{jnt}}}. \tag{8}$$

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:

$$R_{int} = \delta_{RRM,i} + \sum_{k=1}^{K} \sum_{j \neq i} ln(1 + e^{\beta_k (x_{jntk} - x_{intk})}) \tag{9}$$

with $k = 1, ..., K$ is an index across attributes, β_k is a 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 given by:

$$P_{RRM,int} = \frac{e^{-R_{int}}}{\sum_{j=1}^{J} e^{-R_{jnt}}} \tag{10}$$

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). Under decision field theory (DFT), a decision-maker stochastically considers the different attributes of the alternatives over the course of a decision-making process, resulting in the preference values updating iteratively:

$$P_t = S \cdot P_{t-1} + V_t, \tag{11}$$

where P_t is a column vector containing the preference values of each alternative i at time t. S is a feedback matrix with memory and sensitivity parameters and V_t is a valence vector, which determines which attribute is attended to at time t. The valence vector can be described by:

$$V_t = C \cdot M \cdot W_t + \varepsilon_t \tag{12}$$

where C is a contrast matrix used to rescale the values such that they total zero, M is the matrix of attribute values and $W_t = [0..1..0]'$ with entry $k = 1$ if and only if attribute k is the attribute being attended to by the decision-maker at preference updating step t. A DFT model thus estimates a weight, w_k, for the likelihood of attending to attribute k. As the error term, ε_t is drawn from a normal distribution with mean 0 (and a variance which is an estimated parameter), the preference values P_t 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_t (ξ_t and Ω_t, respectively). Hence the probability of choosing alternative j from a set of J alternatives at time t is:

$$P_{DFT}\left[\max_{i \in J} P_t[i] = P_t[j]\right] = \int_{X>0} exp\left[-(X-\Gamma)'\Lambda^{-1}(X-\Gamma)/2\right] / (2\pi|\Lambda|^{0.5}) dX \tag{13}$$

with $X = [P_t[j] - P_t[1], ..., P_t[j] - P_t[J]]'$, $\Gamma = L\xi_t$, $\Lambda = L\Omega_t L'$ 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. (2018a), which incorporates attribute-specific scaling factors [1].

Quantum pairwise comparison (QPCA) Our quantum model is based on the first model (quantum pairwise comparison framework A) defined by Hancock et al. (2019). 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 'state' can then be represented by another vector within this

[1] 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. (2018b) and Hancock et al. (2018a).

space. The action of making a choice is then represented by a projection from this state vector onto the vector representing the chosen alternative (see figures in Hancock et al. (2019). Allowing the state vector to be of unit length results in the set of squared projection lengths onto each of the possible alternatives summing to one. Under QPCA, the length of projection for alternative i (for respondent n in choice task t) is:

$$|\rho_{int}| = \delta_{QPCA,i} + I_0 + \sum_{k=1}^{K} \sum_{j \neq i} wt_{ij} \cdot ln(1 + e^{\beta_k (x_{intk} - x_{jntk})}), \tag{14}$$

where $\delta_{QPCA,i}$ 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. Once these projection lengths have been calculated, the probability for each alternative can be defined simply as:

$$P_{QPCA,jnt} = \frac{|\rho_{jnt}|^2}{\sum_{i=1}^{J} (|\rho_{int}|^2)}, \tag{15}$$

where $i = 1, ...J$ is an index across the possible alternatives.

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

$$U_{int} = \delta_{RAM,i} + \sum_{k=1}^{K} \beta_k \cdot x_{intk} + \sum_{j \neq i} RA(i,j), \tag{16}$$

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

$$RA(i,j) = \frac{A(i,j)}{A(i,j) + D(i,j)}, \tag{17}$$

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

For our SP dataset, we first apply the five different models individually, obtaining the results given in Table 4. We see that DFT obtains the best log-likelihood ahead of QPCA, with the performance of the three logit-style models is poorer and comparatively more similar. As a first step, we look at model averaging across all five models applied to this dataset, where the resulting shares and fit are shown in Table 4. 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 QPCA, 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.

TABLE 4 : 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	QPCA	-3,336.44	6,771	5.70%
5	RAM	-3,354.55	6,791	0.08%
Model averaging		-3,312.40		

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 couple of different rules. We therefore look at the estimation of 15 different latent class structures with two classes per model, thus also allowing for five models where the two classes are of the same type, i.e. looking for taste heterogeneity alone. Table 5 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 would provide 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 5 : Results from latent class models applied to SP dataset

Model	Class 1	Class 2	Log-likelihood	BIC	MA Share
1	MNL	MNL	-3,113.13	6,399	0.0%
2	MNL	RRM	-3,102.66	6,378	0.0%
3	MNL	DFT	-3,099.84	6,380	6.7%
4	MNL	QPC	-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%
7	RRM	DFT	-3,086.79	6,354	11.7%
8	RRM	QPC	-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%
11	DFT	QPC	-3,085.28	6,376	0.0%
12	DFT	RAM	-3,085.38	6,351	0.0%
13	QPC	QPC	-3,095.71	6,380	12.8%
14	QPC	RAM	-3,094.59	6,370	0.0%
15	RAM	RAM	-3,100.27	6,373	0.0%
Model Averaging			-3,071.46		

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 4, 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. Further insights are detailed in Table 6, which for each pair of different decision rules (x, y), gives the difference in model fit between this model and the better fitting model from the latent class models with x in both classes or y in both classes[2]. We see only two cases in favour of decision-rule heterogeneity. The MNL-RRM model outperforms RRM-RRM by 3.67 log-likelihood units (as well as the MNL-MNL model by 10.47 units). Additionally, QPC-RAM has a better log-likelihood than either QPC-QPC or RAM-RAM. However, all other differences are negative, indicating that models with the same decision rule in the 2 different classes frequently perform just as well or better than models with differing decision rules.

Further evidence is given in the model averaging results in Table 5. We see that model averaging obtains a better log-likelihood than any of the individual LC models. Crucially, 81.6% of the share is given to models that each time use just a single decision rule, again highlighting the importance of within-model taste heterogeneity, at least for this data.

TABLE 6 : Differences in log-likelihood between combinations of rules and best fitting model using same rule in both classes

MNL	3.67	-22.05	-11.06	-0.52
	RRM	-9.00	-0.64	-3.95
		DFT	-7.49	-7.59
			QPC	1.12
				RAM

We explore the best example for 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 7. For each model we have coefficients for travel time (TT), log of the fare (LFare[3]), 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 7 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

[2]Note that no formal fit comparisons are made here.

[3]Note that we use a log transform of the fare rather than the fare itself as a cost damping affect is observed.

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 this model are due to heterogeneity in the cost sensitivity rather than heterogeneity in the decision rules. This means that the classes individually have 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 is due to taste rather than decision rule heterogeneity.

TABLE 7 : 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 -3,102.66	-4,431.55	-3,360.43 -3,360.36	-3,363.91
Class LL: Log-likelihood	-3,645.30 -3,102.66	-4,431.55	-3,360.43 -3,360.36	-3,363.91
asc_{alt1}	0.64 (6.42)	0.04 (0.27)	0.39 (5.85)	0.27 (4.17)
asc_{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%

5. CONCLUSIONS

In this paper, we revisit the use of latent class models to capture different 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 taste heterogeneity at the level of a fixed model specification 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 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.

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 for purposes other than taste heterogeneity, 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.

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 survey, 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.

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., 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., and Choudhury, C. F. (2018a). An accumulation of preference: two alternative dynamic models for understanding transport choices. *Submitted*.

Hancock, T. O., Hess, S., and Choudhury, C. F. (2018b). 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., and Choudhury, C. F. (2019). Quantum probability: a new method for modelling travel choices. In *The Transportation Research Board (TRB) 98th Annual Meeting*.

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.

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.