

Bahamonde-Birke, Francisco J.; Ortúzar, Juan de Dios

Working Paper

About the categorization of latent variables in hybrid choice models

DIW Discussion Papers, No. 1527

Provided in Cooperation with:

German Institute for Economic Research (DIW Berlin)

Suggested Citation: Bahamonde-Birke, Francisco J.; Ortúzar, Juan de Dios (2015) : About the categorization of latent variables in hybrid choice models, DIW Discussion Papers, No. 1527, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:

<https://hdl.handle.net/10419/125149>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

1527

Discussion Papers

Deutsches Institut für Wirtschaftsforschung

2015

About the Categorization of Latent Variables in Hybrid Choice Models

Francisco J. Bahamonde-Birke and Juan de Dios Ortúzar

Opinions expressed in this paper are those of the author(s) and do not necessarily reflect views of the institute.

IMPRESSUM

© DIW Berlin, 2015

DIW Berlin
German Institute for Economic Research
Mohrenstr. 58
10117 Berlin

Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
<http://www.diw.de>

ISSN electronic edition 1619-4535

Papers can be downloaded free of charge from the DIW Berlin website:
<http://www.diw.de/discussionpapers>

Discussion Papers of DIW Berlin are indexed in RePEc and SSRN:
<http://ideas.repec.org/s/diw/diwwpp.html>
<http://www.ssrn.com/link/DIW-Berlin-German-Inst-Econ-Res.html>

About the Categorization of Latent Variables in Hybrid Choice Models

Francisco J. Bahamonde-Birke¹

Juan de Dios Ortúzar²

ABSTRACT

Although hybrid choice models are fairly popular nowadays, the way in which different types of latent variables are considered into the utility function has not been extensively analysed. Latent variables accounting for attitudes resemble socio-economic characteristics and, therefore, systematic taste variations and categorizations of the latent variables should be considered. Nevertheless, categorizing a latent variable is not an easy subject, as these variables are not observed and consequently exhibit an intrinsic variability. Under these circumstances it is not possible to assign an individual to a specific group, but only to establish a probability with which an individual should be categorized in given way.

In this paper we explore different ways to categorize individuals based on latent characteristics, focusing on the categorization of latent variables. This approach exhibits as main advantage (over latent-classes for instance) a clear interpretation of the function utilized in the categorization process, as well as taking exogenous information into account. Unfortunately, technical issues (associated with the estimation technique via simulation) arise when attempting a direct categorization.

¹ Energy, Transportation and Environment Department, Deutsches Institut für Wirtschaftsforschung – Berlin
Mohrenstr. 58, 10117-Berlin, Germany, and Technische Universität Berlin
e-mail: bahamondebirke@gmail.com

² Department of Transport Engineering and Logistics, Centre for Urban Sustainable Development (CEDEUS), Pontificia Universidad Católica de Chile, Vicuña Mackenna 4860, Macul, 7820436-Santiago, Chile
e-mail: jos@ing.puc.cl

We propose an alternative to attempt a direct categorization of latent variables (based on an auxiliary variable) and conduct a theoretical and empirical analysis (two case studies), contrasting this alternative with other approaches (latent variable-latent class approach and latent classes with perceptual indicators approach). Based on this analysis, we conclude that the direct categorization is the superior approach, as it offers a consistent treatment of the error term, in accordance with underlying theories, and a better goodness-of-fit.

Keywords: hybrid choice models, latent variables, latent classes, categorization

JEL: C35, C50

1. INTRODUCTION

Although firstly suggested during the 80s (McFadden, 1986; Train *et al.*, 1987), hybrid discrete choice models (HCM) did not become a hot topic in travel behaviour research until revitalized by Ben-Akiva *et al.* (2002). Then, it was possible to solve a series of computational issues that had prevented this approach becoming a standard tool in discrete choice (DC) modelling.

Starting from the usual hypotheses of discrete choice modelling (McFadden, 1974), HCM aim to enrich the model, taking into account unobserved characteristics of the individuals and of the alternatives. This way, indicators (usually stated on a Likert-scale; Likert, 1932) related with attitudes toward life and perceptions about the alternatives on offer, are gathered from a sample of individuals. As these indicators are not considered attributes on their own, but rather an expression of underlying attitudes and perceptions, the modeller usually relies on a Multiple Indicators Multiple Causes (MIMIC) model (Zellner, 1970; Bollen, 1989), allowing for the identification of unobserved latent variables explaining the indicators. It is assumed that these variables also affect the typical utility function of the DC-model, enriching it with exogenous information captured through the indicators.

Although this approach has become popular (v. Acker *et al.*, 2011; Ashok *et al.*, 2002; Daziano and Bolduc, 2013; Bahamonde-Birke and Ortúzar, 2014a; among many others), the way in which different types of latent variables are considered into the utility function has not been extensively analysed; in fact, they are generally introduced in a simplistic additive fashion, while being explained only by characteristics of the individuals (Vredin-Johansson *et al.*, 2006; Yañez *et al.*, 2010; di Ciommo *et al.*, 2013). That being the case, Chorus and Kroesen (2014) argue (rightly) that these types of models do not allow deriving policy implications.

On another hand, Bahamonde-Birke *et al.* (2015) argue that we must distinguished between attitudes (including attitudes toward alternatives and toward perceptions), based on the individual's experience and temperament (Allport, 1935; Pickens, 2005) and real perceptions, related to the way in which individuals perceive their environment (Lindsay and Norman, 1972), as the latter are influenced by the characteristics of the alternatives and not only by those of the individuals (Pickens, 2005). They state that attitudinal latent variables (in contrast with real perceptions) resemble the socioeconomic characteristics of the individuals and hence they should be treated in the same fashion. That being the case, systematic taste variations and categorizations of the latent variables should be considered, rather than just including variables accounting for attitudes in an additive fashion. For example, it may be preferable to identify environmentally aware or wealthy individuals, rather than attempting to associate them with a value in a continuous scale; moreover, it is highly disputable that this characteristics should have a linear impact on decisions.

It is important to note that categorizing a latent variable would depict *de facto* a latent class model (Kamakura and Russell, 1989; Bhat, 1997) - beyond theoretical issues, related to causality - with every category representing a different latent class. In fact, by starting from a latent class model and attempting to enrich the identification of the classes by using indicators, the modeller could face a similar problem.

In this paper we provide a theoretical discussion about the different factors affecting the categorization of latent variables, as well as their consequences. In the same line, we consider previous attempts that have been conducted to deal with this problem. Finally, we propose an alternative way to categorize latent variables and test our hypotheses with the help of two study cases (real data). We compare the results obtained, following the

previously described approaches and offer a discussion about the implications and assumptions of the various methods. This discussion offers valuable insights about the way in which attitudes should be treated in hybrid discrete choice models.

2. METHODOLOGICAL FRAMEWORK

In a HDC framework (Ben-Akiva *et al.*, 2002), individuals are assumed to exhibit utility functions which take the following shape (under the assumption of additive linearity):

$$U_j = \beta_x \cdot X + \beta_\eta \cdot \eta + \varepsilon \quad [2.1],$$

where X stands for observed attributes of the alternatives and characteristics of the individuals, while η is a vector representing the unknown latent variables. β_x and β_η are vectors of parameters to be estimated, and ε an error term. If ε is considered to follow an EV1 distribution with the same mean for all alternatives and a diagonal covariance matrix, the choice probabilities will be given by a Logit model. If the covariance matrix is not diagonal, other member of the Generalized Logit family may describe the choice situation; finally, assuming a Normal distribution leads to the Probit model. For identifiability purposes, some components of the covariance matrix must be constrained without loss of generality (Walker *et al.*, 2007). A given alternative j will be selected if $U_j > U_k \ \forall k \neq j$; in this case the dummy variable y_j would take a positive value.

The latent variables are constructed in accordance with a MIMIC model and are a function of positively observed explanatory variables and, eventually, of other latent constructs (Kamargianni *et al.*, 2014; Link, 2015). This way, assuming a linear additive specification, these structural equations may be described in the following manner:

$$\eta = \alpha_x \cdot X + \alpha_\eta \cdot \eta^* + \nu \quad [2.2],$$

where X , is once more the set of exogenous observed explanatory variables; η and η^* are vectors of latent constructs and ν an error term following any distribution, but usually assumed to be Normal with mean zero and a given covariance matrix Σ_η . Finally, α_x and α_η are matrices of parameters to be estimated.

In this framework, the structural equations set is unidentified¹, and it is mandatory to consider it jointly with a measurement equations set (for identifiability purposes, the utility function itself may be considered as an extra measurement equation; Bahamonde-Birke and Ortúzar, 2015). Assuming a linear specification the latter may be represented as follows:

$$I = \gamma_x \cdot X + \gamma_\eta \cdot \eta + \varsigma \quad [2.3],$$

where I is a vector of exogenously gathered indicators and ς an error term, the distribution of which depends on the assumptions regarding the indicators; when considering the indicators as a continuous output, ς is usually assumed to be Normal (Vredin-Johansson *et al.*, 2006; Daziano and Barla, 2012); in turn, when the indicators are considered to be of discrete nature, ς may be assumed to follow a Logistic distribution (Daly *et al.*, 2012; Hess *et al.*, 2013); in both cases the distributions have zero mean and a diagonal covariance matrix Σ_I ; finally, γ_x and γ_η are matrices of parameters to be estimated.

¹ Even when considering both set of equations jointly, it is still necessary to fix certain parameters without loss of generalization (typically the variances of the structural equations) to achieve identification (Vij and Walker, 2014).

The estimation of the integrated framework may be performed sequentially (estimating first the MIMIC model and then considering the latent variables into the discrete choice component) or simultaneously. Nevertheless, estimating the model sequentially is not advisable as it may lead to biased results (Bahamonde-Birke and Ortúzar, 2014b). For simultaneous estimation, the likelihood function for the integrated framework would take the following form:

$$L = \int_{\eta} P(y | X, \eta; \alpha, \beta, \varepsilon, \nu) \cdot P(I | X, \eta; \alpha, \gamma, \zeta, \nu) \cdot f(\eta | X, \eta^*; \alpha, \nu) \cdot d\eta \quad [2.4],$$

where the first part corresponds to the likelihood of the discrete choice component and the second stands for likelihood of observing the gathered indicators. The third component corresponds to the distribution of latent variables over which the likelihood function must be integrated.

Categorizing a latent variable

This is not an easy subject (in contrast with observed socio-economic characteristics such as income or age), as these variables are not observed and consequently exhibit an intrinsic variability. Under these circumstances, it is not possible to assign an individual to a specific group but only to establish a probability with which an individual should be categorized in given way. Moreover, as latent variables do not exhibit an unequivocal scale, it is not easy to establish thresholds for the categorization, and as such the process appears to be rather arbitrary.

An intuitive approach to deal with the aforementioned problem would be to construct a dummy variable taking a positive value when a given latent variable η_c exceeds a certain threshold ψ . This way, it would be possible to establish a probability with which a certain individual should be categorized in a given way, defining a latent class model indirectly (when the dummy variables are introduced into the utility function). Then, it should be possible to calibrate the threshold and the likelihood function would take the following form (assuming only two categories, but extending the framework for more categories is straightforward):

$$L = \int_{\eta} P(y | X, \eta, 0; \alpha, \beta, \varepsilon, \nu) \cdot P(I | X, \eta; \alpha, \gamma, \zeta, \nu) \cdot f(\eta | X, \eta^*; \alpha, \nu) \cdot d\eta \cdot P_c \\ + \int_{\eta} P(y | X, \eta, 1; \alpha, \beta, \varepsilon, \nu) \cdot P(I | X, \eta; \alpha, \gamma, \zeta, \nu) \cdot f(\eta | X, \eta^*; \alpha, \nu) \cdot d\eta \cdot (1 - P_c) \quad [2.5],$$

with the probability of an individual being categorized in certain fashion (P_c) given by:

$$P_c = P(\eta_c < \psi | s, \alpha, \nu) = P(\alpha_{\eta_c} \cdot X + \alpha_{\eta_c} \cdot \eta^* + \nu_c - \psi < 0) \quad [2.6]$$

Unfortunately HDC models do not exhibit a close form for the probabilities (given their complex error terms structure) and therefore they are estimated using simulation (Ben-Akiva *et al.*, 2002; Bierlaire, 2003). This way we do not observe a continuous distribution for the error term ν of the latent variable η_c , but rather a set of stochastic (or pseudo-stochastic) draws describing a probability function. As a consequence, discontinuity issues

arise, the threshold cannot be calibrated and a perfect convergence of the algorithm cannot be theoretically assured².

a) Latent variable latent class approach (LVLC)

An alternative to overcome the problem above would be to rely on an auxiliary continuous distribution function, to establish the probability with which a certain individual would be categorized in given manner. This approach will lead to the latent variable/latent class specification (Walker and Ben-Akiva; 2002; Hess *et al.*, 2013). This way the likelihood of the individual belonging to a certain latent class is given by the probability of the latent variable being smaller (or larger) than a threshold (to be calibrated), making use of the auxiliary function. If we assume this auxiliary function to be a Logistic distribution, equation [2.6], may be depicted as follows (Hess *et al.*, 2013)³:

$$P_c = \frac{1}{e^{\lambda(\eta_c - \psi)} + 1} = \frac{1}{e^{\lambda(\alpha_{Xc} \cdot X + \alpha_{\eta c} \cdot \eta^* + \nu_c - \psi)} + 1} \quad [2.7],$$

which is indeed a continuous expression. Unfortunately, this approach implies adding (artificially) the variability associated with the auxiliary distribution function (or of the latent class model, if we approach the modelling from this perspective) to the latent variable's own variability ν . Even though it may be argued that the extra variability (which may be calibrated – note that in this specification λ is perfectly identified) represents the error induced in the categorization process, there is no clear statistical justification for that.

To avoid the inclusion of the extra variability, the modeller may exogenously fix λ at a high level, but when doing so mathematical issues arise.

b) Latent class with psychometric indicators approach (LCPI)

This approach (Hurtubia *et al.*, 2014) does not attempt to categorize continuous latent variables *per se*, but rather to include psychometric indicators into a latent class framework. In this structure, it is assumed that the latent variable depends exclusively on positively observed characteristics of the individuals (X)⁴, while its error term follows a Logistic distribution with zero mean and scale parameter λ (which has to be fixed without loss of generality). Under this assumption, the probability of an individual belonging to a certain class is given by:

$$P_c = \frac{1}{e^{\lambda(\alpha_{Xc} \cdot X - \psi)} + 1} \quad [2.8],$$

As can be observed, this approach overcomes the variability issues of the LVLC, considering only the logistically distributed error term. The main difference between this approach and the HDC framework is that in this case the latent variable does not impact directly on the measurement equations, but rather indirectly. This way the measurement equations are just a function of positively observed characteristics of the individuals (X)

² When considering the model sequentially by integrating over the domain of the latent variables, the model achieves perfect convergence, but the threshold remains unidentified and must be fixed a priori (Bahamonde-Birke *et al.*, 2015).

³ In the original specification Hess *et al.* (2013) multiply η_c by a parameter to be estimated and fix λ , but it is straightforward to see that both specifications are equivalent.

⁴ Even though the model by Hurtubia *et al.* (2014) only considers characteristics of the individuals as explanatory variables, it is straightforward to extend it for attributes of the alternatives.

and the parameters to be calibrated depend on the latent class the individual is associated with. Thus, under this approach equation [2.3] exhibits the following structure:

$$I = \gamma_{xc} \cdot X + \varsigma \quad [2.9],$$

where γ_{xc} is a latent class specific matrix of parameters. This leads to the following likelihood function (it is important to note that the original structure does not require integrating over the domain of the categorized latent variable, as the probability functions are given by closed-form expressions; it is straightforward to extend this approach in order to consider non-categorized latent variables):

$$L = P(y | X, 0; \alpha, \beta, \varepsilon, \nu) \cdot P(I | X, 0; \alpha, \gamma, \varsigma, \nu) \cdot P_c \\ + P(y | X, 1; \alpha, \beta, \varepsilon, \nu) \cdot P(I | X, 1; \alpha, \gamma, \varsigma, \nu) \cdot (1 - P_c) \quad [2.10],$$

This approach offers computational advantages (as it requires no integration over the domain of the latent variables) as well as statistical consistency. One disadvantage relies on the fact that it does not allow for the latent variable to be considered directly (continuously) into the measurement equations. This may be necessary, for instance, when using a latent variable accounting for wealth (under the presence of unreported information; Sanko *et al.*, 2014) in a categorized fashion (Bahamonde-Birke and Hanappi, 2015), among other specifications.

Additionally, it must be pointed out that the interpretation of results tends to be rather obscure, especially in the case of the measurement equations, as these may be related to the absence of causality assumptions inherent to latent-class models. Opposite to the HDC framework, where it is assumed that the stated indicators are an expression of underlying attitudes and perceptions, in the LCPI approach the indicators are a tool to improve class-identification (individuals would be more or less likely to belong to a certain latent class given the indicators they have stated) and no assumptions regarding their causes are made. Along the same line, the interpretation of the function used to categorize the individuals remains obscure (which is also a deficiency of latent-class models), as it is not easy to establish a clear meaning for it (opposite to the latent variables, for instance) and often it appears to be rather an ad-hoc function to classify the individuals.

Finally, the complex structure of the LCPI framework (which requires the joint consideration of several sets of latent-class specific equations) is very demanding in terms of data variability. Hence, empirical identification issues arise and simpler structures must be favoured for the categorizing function (at the expense of a theoretical interpretation).

c) Direct categorization of latent variables approach (DCLV)

Starting from the classical HDC framework (equations [2.1], [2.2] and [2.3]), we propose assuming an independent Logistic distribution with zero mean and given scale parameter (which can be fixed without loss of generality) for the error term ν_{cl} of equation [2.2]⁵. Under this assumption equation [2.6] may be represented as a continuous expression.

$$P_c = P(\eta_c < \psi | s, \alpha, \nu) = P(\alpha_{xc} \cdot X + \alpha_{\eta c} \cdot \eta^* + \nu_{cl} - \psi < 0) = \frac{1}{e^{\lambda(\alpha_{xc} \cdot X + \alpha_{\eta c} \cdot \eta^* - \psi)} + 1} \quad [2.11]$$

⁵ Just in the case of the latent variable to be categorized. The error of additional latent variables may follow any distribution.

In order to include the latent variable directly into the measurement equations (as in equation [2.3]), it is still necessary to rely on simulation. For that matter we can construct a second latent variable, which is equivalent to the original one, but includes a simulated error term ν_{c2} , the specification of which is exactly the same as that considered in equation [2.11].⁶ This way the likelihood function is not subject to discontinuity issues, while it still may be integrated over the domain of the latent variables relying on simulation techniques.

It is important to note that the DCLV framework would dominate the LCPI approach, as in this case is possible to include class specific parameters into the measurement equations (via the inclusion of a dummy associated with the categorized latent variable). Nevertheless, it must be pointed out that the inclusion of latent class specific parameters must be carefully considered and be in accordance with any underlying hypotheses. In other case, the modelling would be subject to the same problems of the latent class structures (LC, LCPI) and, therefore, it may be advisable to omit using latent class specific parameters in the measurement equation.

Thus, the general specification (considering latent class specific parameters in the measurement equations) for the likelihood function following the DCLV approach (considering only two categories) is the following:

$$L = \int_{\eta} P(y | X, \eta, 0; \alpha, \beta, \varepsilon, \nu) \cdot P(I | X, \eta, 0; \alpha, \gamma, \zeta, \nu) \cdot f(\eta | X, \eta^*; \alpha, \nu) \cdot d\eta \cdot P_c \quad [2.12],$$

$$+ \int_{\eta} P(y | X, \eta, 1; \alpha, \beta, \varepsilon, \nu) \cdot P(I | X, \eta, 1; \alpha, \gamma, \zeta, \nu) \cdot f(\eta | X, \eta^*; \alpha, \nu) \cdot d\eta \cdot (1 - P_c)$$

Limitations

The main limitation associated with all the approaches mentioned above, is the necessity of a continuous closed-form expression (as the Logistic distribution) for the categorizing function. This may appear to be a rather innocuous limitation, but it has major implications when working with panel data (or pseudo-panel data). In this case, integration over the domain of the latent variables should be performed on an individual level (e.g. via random panel effects; Bhat and Gossen, 2004). Although it is possible to include random panel effects, the fact that the error terms associated with the categorizing function are independent, would cause the categorization to be, at least to a certain extent, independent for all choices of the same individual.

Another inconvenient is related with the non-monotonicity of the categorizing function. This leads to the existence of local optima and therefore the categorization's threshold will depend on the starting value. Nevertheless, this problem is common to all latent class approaches.

3. CASE STUDIES

To analyse the different approaches presented in Section 2, we compared them making use of empirical data. Even though in the previous section we allowed for the DCLV framework to consider latent class specific parameters in the measurement equations, it is highly disputable if it would be advisable (given the lack theoretical justification);

⁶ It is important to note that it would be necessary only to fix the scale parameter of one error term (the continuous or the simulated one) and the second could be estimated; but this is not advisable for statistical and theoretical consistency (we are assuming that both variables represent the same error term).

moreover HDC models usually do not even consider positively observed attributes in the measurement equations. For that reason, in our analysis we will just consider latent variables as explanatory variables in the measurement equations, when following the LVLC and DCLV approaches (the structure of both approaches will only differ in the specification of the error terms). This allows for a direct comparison between both approaches.

LCPI models exhibit a totally different structure and, therefore, a fair comparison is not possible. First, because of the aforementioned empirical identification issues, it would not be possible to consider the same degree of complexity for the categorizing function (structural equations). In fact, when doing so, it was only possible to work with one of five indicators in Case study 1 and with none (out of two) in Case study 2, and for that reason simpler specifications should be analysed at the expense of the meaning of the classes being constructed. Second, a comparison of the goodness-of-fit would be spurious, as the LCPI must necessarily be more likely than models estimated according to the LVLC or DCLV specifications. In the LCPI classes are defined exclusively to maximize the goodness-of-fit and are not based on theoretical assumptions (additionally, the measurement equations would have more degrees-of-freedom). As a matter of fact, if we consider the likelihood of the discrete choice component only, it is clear that a simple latent class model (with no indicators) would outperform every competing approach (including the LCPI; Hurtubia *et al.*, 2014), as in this case the latent classes do not have to satisfy, additionally, a distribution of indicators or theoretical assumptions.

As a consequence we do not consider the LCPI in our empirical analysis, and our conclusions in this sense will be based exclusively on the aforementioned theoretical arguments.

Case study 1 - Electromobility in Austria

Our first case study comes from the DEFINE project and a web-based survey conducted in Austria during February 2013. The survey considered a SP-experiment and was set in the context of choosing between different options for electromobility (Bahamonde-Birke and Hanappi, 2015). The sample was representative of the Austrian society and consisted of 1,449 respondents, where 787 of them were considered in the discrete choice experiment.

This vehicle purchase experiment used a labelled experimental design, including four alternatives referring to different propulsion technologies: conventional vehicles (CV), plug-in hybrid-electric vehicles (PHEV), hybrid-electric vehicles (HEV) and electric vehicles (EV). Each alternative was described by the following attributes: purchase price (PP), power (PS), fuel costs (FC) and maintenance costs (MC). In addition to these attributes, the EV was further characterized by the following attributes: full driving range (RA), availability of charging stations (LS) and policy incentives (IM). Charging station availability varied across three categories (low, intermediate and high) and was described qualitatively using a separate pop-up box. Policy incentives included a Park and Ride subscription for one year (IM2), investment subsidies to support private charging stations (IM3), and a one-year-ticket for public transport (IM4).

Additionally, attitudinal indicators related to the degree of agreement with eight different sentences were collected. Bahamonde-Birke and Hanappi (2015) considered the following five of them to construct a latent variable related to the environmental attitude of the individuals: *“I am an ecologically aware person”* (EcAwareness), *“I pay attention to regional origins when shopping foods and groceries”* (LocalFood), *“I buy ecologically*

friendly products” (EcoFriendly), “*Environmental protection measures should be enacted even if they result in job losses*” (Protection) and “*I pay attention to the CO2 footprint of the products I buy*” (CO2Footprint). The level of agreement was stated using a six points Likert scale (Likert, 1932). Table 1 presents an overview of the variables that are relevant to our study⁷.

Table 1 – Definition of the variables considered in the model. Case Study 1

<i>Variable</i>	<i>Definition</i>
<i>MidSkill</i>	Dummy variable indicating a career and technical education.
<i>HighSkill</i>	Dummy variable indicating a college education or higher.
<i>Vienna</i>	Dummy variable indicating a residence in Vienna.
<i>Male</i>	Dummy variable indicating masculine gender.
<i>Old</i>	Dummy variable indicating individuals older than 60 years
<i>MidAge</i>	Dummy variable indicating individuals older than 35 years, but no older than 60 year.
<i>Carsharing</i>	Dummy variable indicating that the individual relies on Car Sharing on a regular basis.
<i>CarUser</i>	Dummy variable indicating that the individual drives to their main occupational activity on a regular basis.
<i>PP</i>	Purchase price in €.
<i>FC, MC</i>	Fuel and maintenance cost in € / 100 km., respectively.
<i>PS</i>	Power of the engine in hp.
<i>RA</i>	Driving range in km.
<i>IM2, IM3, IM4</i>	Dummy variables indicating the execution of the respective policy incentive.
<i>LSMid, LSHigh</i>	Dummy variables indicating medium or high availability of loading stations for EV.
<i>LV Green</i>	Latent variable Green accounting for environmental awareness.
<i>CLV Green</i>	Categorized latent variable Green accounting for individuals with high environmental awareness.
<i>EcAwareness</i>	Attitudinal Indicator for “I am an ecologically aware person”.
<i>LocalFood</i>	Attitudinal Indicator for “I pay attention to regional origins when shopping foods and groceries”.
<i>EcoFriendly</i>	Attitudinal Indicator for “I buy ecologically friendly products”.
<i>Protection</i>	Attitudinal Indicator for “Environmental protection measures should be enacted even if they result in job losses”.
<i>CO2Footprint</i>	Attitudinal Indicator for “I pay attention to the CO2 footprint of the products I buy”.

We estimated two models following the approaches presented in the previous section, using PythonBiogeme (Bierlaire, 2003). Taking advantage of having just one latent variable, the log-likelihood was computed using numerical integration. The latent variable (*LV Green*) accounts for environmental awareness and was categorized in two levels. The inclusion of an individual specific dummy variable (accounting for environmentally concerned individuals) associated with the alternatives HEV, PHEV and BEV, was the only difference between both classes. The results for the models are presented in Table 2. The results of the t-test for statistical significance are presented in parenthesis and the log-likelihood for the overall model, as well as for the DC component only, are also reported (the results for the measurement equations are presented in the appendix). The indicators were considered discrete outputs following an ordered logit (OLM) approach (Bahamonde-Birke and Ortúzar, 2015).

⁷ Bahamonde-Birke and Hanappi (2015) considered further variables and also estimated more involved models than those considered in this study. For the purposes of this work, this specification was considered appropriate, as additional complexity would only add noise to our analysis, as well as computational complexity.

Table 2 – Parameter estimates. Case study 1.

<i>Variable</i>	<i>Equation</i>	<i>LVLC</i>		<i>DCLV</i>	
<i>Vienna</i>	<i>S.E. LV Green</i>	-0.333	(-2.56)	-0.182	(-1.71)
<i>Male</i>	<i>S.E. LV Green</i>	-0.597	(-5.58)	-0.587	(-5.74)
<i>HighSkill</i>	<i>S.E. LV Green</i>	1.05	(6.16)	0.879	(5.89)
<i>MidSkill</i>	<i>S.E. LV Green</i>	0.536	(3.85)	0.477	(3.72)
<i>Old</i>	<i>S.E. LV Green</i>	1.09	(7.18)	0.955	(6.67)
<i>MidAge</i>	<i>S.E. LV Green</i>	0.657	(5.15)	0.615	(5.14)
<i>Carsharing</i>	<i>S.E. LV Green</i>	1.33	(4.9)	0.988	(4.99)
<i>CarUser</i>	<i>S.E. LV Green</i>	-0.567	(-5.53)	-0.474	(-5.7)
<i>Threshold</i>	<i>C.F</i>	3.76	(5.27)	2.23	(7.18)
<i>Scale Parameter</i>	<i>C.F</i>	0.538	(5.06)	1	(fixed)
<i>ASC_CV</i>	<i>Utility CV</i>	0	(fixed)	0	(fixed)
<i>ASC_HEV</i>	<i>Utility HEV</i>	-0.0165	(-0.07)	0.0945	(0.41)
<i>ASC_PHEV</i>	<i>Utility PHEV</i>	-0.235	(-0.94)	-0.207	(-0.88)
<i>ASC_EV</i>	<i>Utility EV</i>	-1.09	(-2.65)	-0.982	(-2.36)
<i>PP</i>	<i>Utility CV</i>	-1.21	(-7.92)	-1.27	(-9.07)
<i>PP</i>	<i>Utility HEV</i>	-1.8	(-22.39)	-1.79	(-22.85)
<i>PP</i>	<i>Utility PHEV</i>	-1.76	(-19.94)	-1.77	(-20.48)
<i>PP</i>	<i>Utility EV</i>	-1.66	(-9.79)	-1.73	(-10.69)
<i>MC</i>	<i>Utility CV, HEV, PHEV, EV</i>	-21.1	(-9.32)	-20.6	(-9.6)
<i>FC</i>	<i>Utility CV, HEV, PHEV, EV</i>	-21.4	(-14.96)	-20.7	(-15.87)
<i>PS</i>	<i>Utility CV</i>	0.0327	(5.22)	0.0341	(5.74)
<i>PS</i>	<i>Utility HEV</i>	0.0353	(7.18)	0.0348	(7.37)
<i>PS</i>	<i>Utility PHEV</i>	0.0354	(7.2)	0.0367	(7.5)
<i>PS</i>	<i>Utility EV</i>	-0.000269	(-0.05)	-0.000232	(-0.05)
<i>PS * Male</i>	<i>Utility CV</i>	-0.019	(-4.14)	-0.0195	(-4.21)
<i>PS * Male</i>	<i>Utility HEV</i>	-0.0164	(-3.38)	-0.0164	(-3.38)
<i>PS * Male</i>	<i>Utility PHEV</i>	-0.0155	(-3.16)	-0.0162	(-3.26)
<i>PS * Male</i>	<i>Utility EV</i>	-0.00302	(-0.55)	0.000491	(0.09)
<i>MidAge</i>	<i>Utility HEV</i>	-0.266	(-2.5)	-0.287	(-2.65)
<i>MidAge</i>	<i>Utility PHEV</i>	-0.383	(-3.46)	-0.374	(-3.48)
<i>MidAge</i>	<i>Utility EV</i>	-0.784	(-4.55)	-0.946	(-4.91)
<i>Old</i>	<i>Utility HEV</i>	-1.18	(-5.82)	-1.19	(-5.63)
<i>Old</i>	<i>Utility PHEV</i>	-1.44	(-6.24)	-1.35	(-6.07)
<i>Old</i>	<i>Utility EV</i>	-2.55	(-9.02)	-2.88	(-8.77)
<i>CLV Green</i>	<i>Utility HEV</i>	3.48	(6.86)	3.19	(6.56)
<i>CLV Green</i>	<i>Utility PHEV</i>	3.25	(6.08)	2.71	(4.84)
<i>CLV Green</i>	<i>Utility EV</i>	5.77	(8.5)	5.89	(9.83)
<i>RA</i>	<i>Utility EV</i>	0.00444	(6.25)	0.00459	(6.81)
<i>LSMid</i>	<i>Utility EV</i>	0.0121	(0.07)	0.0426	(0.25)
<i>LSHigh</i>	<i>Utility EV</i>	0.661	(4.22)	0.71	(4.43)
<i>IM3</i>	<i>Utility EV</i>	0.306	(2.19)	0.305	(2.14)
<i>Log-likelihood</i>		-16,177.759		-16,169.571	
<i>Overall Model</i>					
<i>Log-likelihood</i>		-6,612.876		-6,605.847	
<i>DC Component</i>					

As can be observed, the differences between the parameters obtained following both approaches are not large and many are not statistically different. This is valid for all parameters of the structural and measurement equations as well as for the utility functions, with the exception of the threshold parameter of the categorizing function. This parameter evidently exhibits a different value, as the LVLC approach is associated with a greater error and a wider distribution, implying that the threshold must be located further away from the expected value, in order to capture the same individuals.

Regarding goodness-of-fit, the DCLV approach exhibits a better performance than the LVLC. This superior adjustment is mostly explained by the discrete choice component. This result is in accordance with our expectations, as the LVLC considers an additional error component, which is indeed unnecessary for the estimation of the model. Along this

line, the fact that the improvement in goodness-of-fit is mostly explained by the discrete choice component, is based on the fact that the categorization (and its additional error term) affects only the utility functions

Case study 2 – Modal choice in Germany

In this SP experiment respondents were asked to choose between different interurban public transport alternatives in Germany (regional and intercity trains, and interurban coaches). The experiment was carried out in three waves (January 2014, March 2014 and April/May 2014), contacting students and employees of two universities in Berlin (the Technische Universität Berlin and the Humboldt-Universität zu Berlin), as well as employees of member institutions of the Leibniz-Gemeinschaft (for further details refer to Bahamonde-Birke *et al.*, 2014). Respondents were required to choose between a first pivotal alternative, representing a trip previously described, and a new one. Alternatives were described in terms of their travel time, fare, number of transfers, mode of transport - regional trains (RE), intercity trains (FVZ) and coaches (LB) - and a safety level.

The original study considered several indicators, associated with different and complex latent variables. For the purposes of this work we only considered one latent variable (TrainFan) again, which is associated with the following two indicators: “*Investing on the development of high-speed trains should be encouraged*” (HSTrains) and “*New high-speed rail lines should be built*” (RailLines). Originally the indicators were stated in a 10-points Likert scale. For computational issues in this application (opposite to the original study), we reduced it to only five, aggregating consecutive levels. Table 3 presents an overview of the variables considered.

Table 3 – Definition of the variables considered in the model. Case Study 2

<i>Variable</i>	<i>Definition</i>
<i>Old</i>	Dummy variable indicating individuals older than 50 years.
<i>Bahncard</i>	Dummy variable indicating ownership of a Deutsche Bahn yearly discount card.
<i>Woman</i>	Dummy variable indicating feminine gender.
<i>VeryLowIncome</i>	Dummy variable indicating a net income under 700€ p.m.
<i>LowIncome</i>	Dummy variable indicating a net income between 700€ and 1,500€ p.m.
<i>MiddleIncome</i>	Dummy variable indicating a net income between 1,500€ and 2,500€ p.m.
<i>HighIncome</i>	Dummy variable indicating a net income over 2,500€ p.m.
<i>Price</i>	Travel fare in €.
<i>TravelTime</i>	Travel time in minutes.
<i>Transfers</i>	Number of transfers.
<i>SafetyLevel</i>	Number of severely injured passengers and the number of fatalities in the overall network over a year.
<i>LV TrainFan</i>	Latent variable TrainFan.
<i>HSTrains</i>	Attitudinal Indicator for “Investing on the development of high-speed trains should be encouraged”.
<i>RailLines</i>	Attitudinal Indicator for “New high-speed rail lines should be built”.

Again, models were estimated following the LVLC and DCLV approaches. Estimation was performed using PythonBiogeme and considering numerical integration for the computation of the likelihood function. Indicators were considered discrete outcomes (OLM). Table 4 presents the results for both models. The structure of the table is the same as in the previous case (results for the measurement equations are presented, again, in the appendix).

Table 4 – Parameter estimates. Case study 2.

<i>Variable</i>	<i>Equation</i>	<i>MCT2</i>		<i>OLM2</i>	
<i>Old</i>	<i>S.E. LV TrainFan</i>	0.207	-0.36	-0.093	-0.41
<i>Bahncard</i>	<i>S.E. LV TrainFan</i>	0.684	-5.41	0.673	-5.46
<i>Woman</i>	<i>S.E. LV TrainFan</i>	-0.686	(-5.64)	-0.663	(-5.75)
<i>MiddleIncome</i>	<i>S.E. LV TrainFan</i>	0.368	-3.19	0.426	-3.24
<i>HighIncome</i>	<i>S.E. LV TrainFan</i>	0.515	-2.53	0.606	-2.82
<i>Threshold</i>	<i>C.F</i>	0.622	(1.44)	-0.346	(-0.59)
<i>Scale Parameter</i>	<i>C.F</i>	10	(fixed)	1	(fixed)
<i>Inertia</i>	<i>Utility Alternative 1</i>	0.336	-13.48	0.345	-13.48
<i>FVZ</i>	<i>Utility Alternative 1 and 2</i>	0	(fixed)	0	(fixed)
<i>LB</i>	<i>Utility Alternative 1 and 2</i>	-1.17	(-12.65)	-0.559	(-12.62)
<i>RE</i>	<i>Utility Alternative 1 and 2</i>	-0.167	(-9.59)	0.448	(-9.52)
<i>Travel Time</i>	<i>Utility Alternative 1 and 2</i>	-1.49	(-26.57)	-1.56	(-26.57)
<i>Ln(Price) * Very Low Income</i>	<i>Utility Alternative 1 and 2</i>	-5.29	(-31.91)	-5.41	(-31.9)
<i>Ln(Price) * Low Income</i>	<i>Utility Alternative 1 and 2</i>	-4.7	(-26.28)	-4.85	(-26.26)
<i>Ln(Price) * Middle Income</i>	<i>Utility Alternative 1 and 2</i>	-3.63	(-14.98)	-3.8	(-15)
<i>Ln(Price) * High Income</i>	<i>Utility Alternative 1 and 2</i>	-2.54	(-7.81)	-2.67	(-7.84)
<i>Safety Level</i>	<i>Utility Alternative 1 and 2</i>	-0.00374	(-4.59)	-0.00408	(-4.59)
<i>Transfers</i>	<i>Utility Alternative 1 and 2</i>	-0.446	(-18.25)	-0.46	(-18.25)
<i>FVZ * CLV TrainFan</i>	<i>Utility Alternative 1 and 2</i>	0.716	(4.93)	1.42	(5.27)
<i>Log-likelihood</i>		-10,543.642		-10,535.416	
<i>Overall Model</i>					
<i>Log-likelihood</i>		-7,461.029		-7,453.871	
<i>DC Component</i>					

While in Case 1 all estimators (aside from the threshold) were not statistically different in both models, in Case 2 the parameters associated with the categorized latent variable are statistically different. As the categorized latent variable is considered in conjunction with the modal parameter of the intercity trains, it also affects the remaining modal parameters. This difference may be attributed to the fact that changing the distribution (the variability) of the categorizing function may allow for identifying different groups of people. In fact, taking a look at the threshold parameters, it seems (accounting for the wider variability of the LVLC) that the thresholds have been set at a different level.

Notoriously in this case, when following the LVLC approach, the parameter associated with the scale parameter λ of the categorizing function diverged (at least in computational terms). For that reason it was necessary to fix it as 10 (greater values would lead to computational issues, when computing the Fisher information matrix, but neither the estimates nor the final value of the log-likelihood would be majorly affected).

Again, the DCLV approach offers a better adjustment than the LVLC model, which may be mainly related to a better explanation of the discrete choices. Interestingly in this case, the scale parameter of the LVLC model suggests that the categorizing function should be as similar to a Dirac delta function as possible. In this limit, the LVLC model would collapse to the DCLV. Nevertheless, achieving this limit (or getting close to it) is not possible due to computational limitations; therefore, it is impossible to get rid of the error induced through the LVLC approach.

4. CONCLUSIONS

Categorizing a latent variable or in general terms, defining the conditions to categorize individuals based on their latent characteristics should offer significant advantages in discrete choice modelling. Nowadays, the dominant approach for categorizing individuals is the latent class approach, but this method may be subject to criticism given its inherent lack of causality assumptions and the obscure interpretation of the functions used to define

class-membership. Additionally, the approach does not take advantage of additional information, such as the perceptual and attitudinal indicators. This criticism may be overruled when latent-classes are used for identifying objective properties (such as missing information, lexicographic respondents, etc.); however, when attempting to model attitudes, perceptions, values and other latent characteristics of the individuals, alternative approaches may be favoured.

In this work we explore different ways to improve the categorization of individuals, focusing on the categorization of latent variables. This approach exhibits, as main advantage, a clear interpretation of the function used in the categorization process (the latent variable), as well as taking exogenous information (perceptual and attitudinal indicators) into account. Unfortunately, technical issues (associated with the estimation technique via simulation) arise when attempting a direct categorization. Therefore, alternative strategies have been proposed.

First, we considered the LVLC approach, which effectively overcomes these technical issues, but is associated with an artificial increase in the error. A second method is the LCPI approach, but it is guilty of the same deficiencies of latent class models: lack of causality assumptions, obscure interpretation of the class-membership function and additionally in this case, obscure interpretation of the measurement equations for the indicators and major issues related to empirical identifiability. It is important to mention that, by definition, this approach should offer superior goodness-of-fit than alternatives approaches, as the class-membership function is constructed to maximize the adjustment and not in order to satisfy *a priori* theoretical hypotheses; we consider this one of its major disadvantages. Finally, we propose an alternative way to attempt a direct categorization of latent variables (DCLV). This approach overcomes the error issues of the LVLC.

We tested the aforementioned approaches with the help of two case studies. The LCPI approach was not considered (based on the just described theoretical concerns, as well as its empirical identifiability issues). Hence, we just considered the LVLC and DCLV approaches.

In the first case study, it was not possible to establish the existence of major differences in the estimated parameters (aside from the threshold parameter), but in the second case study, the differences in the variability of the class-membership function led to a different categorization. In line with our expectations, the DCLV approach offered a superior goodness-of-fit, as the LVLC introduced an additional error term. The improvements in goodness-of-fit were mainly explained by a better adjustment of the discrete choice component.

Based on our empirical results and theoretical analysis, the DCLV appears to be the superior approach. On the one side, it offers a consistent treatment of the error term, which is in accordance with the underlying theory. Along this line, it offers a better performance than the LVLC approach, in terms of the goodness-of-fit. Finally, the approach performed stably (no identification issues) for the considered case studies (opposite to the LCPI).

To end, it must be remarked that all approaches are based on non-monotonic categorizing functions leading to the existence of local optima. Therefore, different starting values must be evaluated. Additionally, it must be mentioned that the approaches do not allow considering the correlation among the responses provided by the same individual, when working with panel-data (at least to a certain extent). Further research must be conducted in this regard.

ACKNOWLEDGMENTS

This paper is partially based on scientific work done during the DEFINE (Development of an Evaluation Framework for the Introduction of Electromobility) project. We gratefully acknowledge the funding for DEFINE (<https://www.ihs.ac.at/projects/define/>) as part of the ERA-NET Plus Electromobility+ call by the EU-Commission and national funding institutions: the Ministry for Transport, Innovation and Technology (Austria), the Federal Ministry of Transport and Digital Infrastructure, formerly Federal Ministry for Transport, Building and Urban Development (Germany), and the National Centre for Research and Development (Poland). We are also grateful to the Institute in Complex Engineering Systems (ICM: P05-004F; FONDECYT: FB016), the All Latitudes and Cultures BRT Centre of Excellence funded by the Volvo Research and Educational Foundations, the Alexander von Humboldt Foundation and the Centre for Sustainable Urban Development, CEDEUS (Conicyt/Fondap/15110020). The authors would also like to thank Prof. Michel Bierlaire for his useful comments and insights. All errors are the authors' sole responsibility

REFERENCES

- van Acker, V., Mokhtarian, P.L. and Witlox, F. (2011). Going soft: on how subjective variables explain modal choices for leisure travel. *European Journal of Transport and Infrastructure Research* **11**, 115-147.
- Allport, G.W. (1935). Attitudes. In Murchison, C. (ed.). *A Handbook of Social Psychology*. Clark University Press, Worcester, MA.
- Ashok, K., Dillon, W. and Yuan, S. (2002). Extending discrete choice models to incorporate attitudinal and other latent variables. *Journal of Marketing Research* **39**, 31-46.
- Bahamonde-Birke, F.J. and Hanappi, T. (2015). The potential of electromobility in Austria: an analysis based on hybrid choice models. *Discussion paper* **1472**, DIW-Berlin.
- Bahamonde-Birke, F., Kunert, U., Link, H. and Ortúzar, J. de D. (2014). Bewertung der angebotsmerkmale des personenfernverkehrs vor dem hintergrund der liberalisierung des fernbusmarktes. *Zeitschrift für Verkehrswissenschaft* **85**, 107-123 (in German).
- Bahamonde-Birke, F.J., Kunert, U., Link, H. and Ortúzar, J. de D. (2015). About attitudes and perceptions - finding the proper way to consider latent variables in discrete choice models. *Discussion paper* **1474**, DIW-Berlin.
- Bahamonde-Birke, F.J. and Ortúzar, J. de D. (2014a). On the variability of hybrid discrete choice models. *Transportmetrica* **10A**, 74-88.
- Bahamonde-Birke, F.J. and Ortúzar, J. de D. (2014b). Is sequential estimation a suitable second best for estimation of hybrid choice models? *Transportation Research Record* **2429**, 51-58.
- Bahamonde-Birke, F.J. and Ortúzar, J. de D. (2015). Analyzing the continuity of attitudinal and perceptual indicators in hybrid choice models. *14th International Conference on Travel Behaviour Research (IATBR)*, Windsor, U.K.

- Bhat, C.R. (1997). An endogenous segmentation mode choice model with an application to intercity travel. *Transportation science* **31**, 34-48.
- Bhat, C.R. and Gossen, R. (2004). A mixed multinomial logit model analysis of weekend recreational episode type choice. *Transportation Research* **38B**, 767-787.
- Ben-Akiva, M.E., Walker, J.L., Bernardino, A.T., Gopinath, D.A., Morikawa, T. and Polydoropoulou, A. (2002). Integration of choice and latent variable models. In H.S. Mahmassani (ed.), *In Perpetual Motion: Travel Behaviour Research Opportunities and Challenges*. Pergamon, Amsterdam.
- Bierlaire, M. (2003). BIOGEME: a free package for the estimation of discrete choice models, *3rd Swiss Transportation Research Conference*, Ascona, Switzerland.
- Bollen, K.A. (1989). *Structural Equations with Latent Variables*. John Wiley and Sons, Chichester.
- Chorus, C. and Kroesen, M. (2014). On the (im-)possibility of deriving transport policy implications from hybrid choice models. *Transport Policy* **36**, 217-222.
- di Ciommo, F., Monzon, A. and Fernandez-Heredia, A. (2013). Improving the analysis of road pricing acceptability surveys by using hybrid models. *Transportation Research* **49A**, 302-316.
- Daly, A., Hess, S., Patruni, B., Potoglou, D. and Rohr, C. (2012). Using ordered attitudinal indicators in a latent variable choice model: a study of the impact of security on rail travel behaviour. *Transportation* **39**, 267-297.
- Daziano, R.A. and Barla, P. (2012). On the way to sustainable commuting to campus: stated, revealed and attitudinal response using a mixed logit model. *Kuhmo Nectar Conference of ITEA 2012*, Berlin, Germany.
- Daziano, R.A. and Bolduc, D. (2013). Incorporating pro-environmental preferences towards green automobile technologies through a Bayesian hybrid choice model. *Transportmetrica* **9A**, 74-106.
- Hess, S., Shires, J. and Jopson, A. (2013). Accommodating underlying pro-environmental attitudes in a rail travel context: application of a latent variable latent class specification. *Transportation Research* **25D**, 42-48.
- Hurtubia, R., Nguyen, M.H., Glerum, A. and Bierlaire, M. (2014). Integrating psychometric indicators in latent class choice models. *Transportation Research* **64A**, 135-146.
- Kamakura, W.A. and Russell, G. (1989). A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research* **26**, 379-390.
- Kamargianni, M., Ben-Akiva, M. and Polydoropoulou, A. (2014). Integrating social interaction into hybrid choice models. *Transportation* **41**, 1263-1285.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology* **140**, 1-55.
- Lindsay, P.H. and Norman, D.A. (1972). *Human Information Processing: An Introduction to Psychology*. Academic Press, New York.
- Link, H. (2015). Is car drivers' response to congestion charging schemes based on the correct perception of price signals? *Transportation Research* **71A**, 96-109.

- McFadden, D. (1974). Conditional logit analysis of qualitative choice behaviour. In P. Zarembka (ed.), *Frontiers in Econometrics*, 105-142. Academic Press, New York.
- McFadden, D. (1986). The choice theory approach to market research. *Marketing Science*, **5**, 275-297.
- Pickens, J. (2005). *Organizational Behavior in Health Care*. Jones and Bartlett Publishers, Sudbury, MA.
- Sanko, N., Hess, S., Dumont, J. and Daly, A. (2014). Contrasting imputation with a latent variable approach to dealing with missing income in choice models. *Journal of Choice Modelling* **12**, 47-57.
- Train, K.E., McFadden, D.L. and Goett, A.A. (1987). Consumer attitudes and voluntary rate schedules for public utilities. *Review of Economics and Statistics* **64**, 383-91.
- Vij, A. and Walker, J. (2014). Hybrid choice models: the identification problem. In S. Hess and A. Daly (eds.), *Handbook of Choice Modelling*, Edward Elgar Publishing, Cheltenham.
- Vredin-Johansson, M., Heldt, T. and Johansson, P. (2006). The effects of attitudes and personality traits on mode choice. *Transportation Research* **40A**, 507-525.
- Yáñez M.F., Raveau, S. and Ortúzar, J. de D. (2010) Inclusion of latent variables in mixed logit models: modelling and forecasting. *Transportation Research* **44A**, 744-753.
- Walker, J. and Ben-Akiva, M. (2002). Generalized random utility model. *Mathematical Social Sciences* **43**, 303-343.
- Walker, J.L., Ben-Akiva, M. and Bolduc, D. (2007). Identification of parameters in normal error component logit-mixture (NECLM) models. *Journal of Applied Econometrics* **22**, 1095-1125.
- Zellner A. (1970). Estimation of regression relationships containing unobservable variables. *International Economic Review* **11**, 441-454.

Appendix 1

Appendix 1 – Parameter estimates for the measurement equations. Case study 1.

<i>Variable</i>	<i>Equation</i>	<i>LVLC</i>		<i>DCLV</i>	
<i>Scale Parameter</i>	<i>M.E. EcAwareness</i>	0.959	(16.14)	0.954	(16.1)
<i>Threshold 1</i>	<i>M.E. EcAwareness</i>	-3.25	(-15.18)	-3.18	(-15.29)
<i>Threshold 2</i>	<i>M.E. EcAwareness</i>	0.0152	(0.09)	0.0719	(0.42)
<i>Threshold 3</i>	<i>M.E. EcAwareness</i>	3.05	(12.85)	3.1	(13.22)
<i>Threshold 4</i>	<i>M.E. EcAwareness</i>	5.06	(14.74)	5.1	(14.9)
<i>Threshold 5</i>	<i>M.E. EcAwareness</i>	7.01	(12.11)	7.05	(12.16)
<i>Scale Parameter</i>	<i>M.E. LocalFood</i>	1.18	(14.89)	1.22	(14.89)
<i>Threshold 1</i>	<i>M.E. LocalFood</i>	-1.76	(-9.58)	-1.68	(-9.51)
<i>Threshold 2</i>	<i>M.E. LocalFood</i>	0.655	(3.74)	0.692	(4.11)
<i>Threshold 3</i>	<i>M.E. LocalFood</i>	2.59	(12.3)	2.59	(12.79)
<i>Threshold 4</i>	<i>M.E. LocalFood</i>	4.23	(15.2)	4.2	(15.63)
<i>Threshold 5</i>	<i>M.E. LocalFood</i>	7.17	(10.56)	7.09	(10.7)
<i>Scale Parameter</i>	<i>M.E. EcoFriendly</i>	0.988	(15.94)	1.02	(16.06)
<i>Threshold 1</i>	<i>M.E. EcoFriendly</i>	-2.68	(-12.98)	-2.58	(-13.01)
<i>Threshold 2</i>	<i>M.E. EcoFriendly</i>	-0.727	(-4.12)	-0.664	(-3.92)
<i>Threshold 3</i>	<i>M.E. EcoFriendly</i>	1.28	(6.9)	1.3	(7.32)
<i>Threshold 4</i>	<i>M.E. EcoFriendly</i>	2.88	(12.85)	2.88	(13.32)
<i>Threshold 5</i>	<i>M.E. EcoFriendly</i>	6.77	(12.65)	6.69	(12.87)
<i>Scale Parameter</i>	<i>M.E. Protection</i>	0.442	(13.06)	0.446	(13.05)
<i>Threshold 1</i>	<i>M.E. Protection</i>	-7.7	(-13.42)	-7.59	(-13.39)
<i>Threshold 2</i>	<i>M.E. Protection</i>	-3.33	(-11.57)	-3.25	(-11.54)
<i>Threshold 3</i>	<i>M.E. Protection</i>	1.3	(5.53)	1.34	(5.83)
<i>Threshold 4</i>	<i>M.E. Protection</i>	4.12	(10.97)	4.14	(11.15)
<i>Threshold 5</i>	<i>M.E. Protection</i>	8.04	(11.84)	8.03	(11.91)
<i>Scale Parameter</i>	<i>M.E. CO2Footprint</i>	0.951	(16.84)	0.95	(16.77)
<i>Threshold 1</i>	<i>M.E. CO2Footprint</i>	-4.39	(-17.52)	-4.32	(-17.61)
<i>Threshold 2</i>	<i>M.E. CO2Footprint</i>	-1.92	(-10.27)	-1.86	(-10.29)
<i>Threshold 3</i>	<i>M.E. CO2Footprint</i>	0.276	(1.56)	0.329	(1.92)
<i>Threshold 4</i>	<i>M.E. CO2Footprint</i>	1.89	(9.4)	1.93	(9.85)
<i>Threshold 5</i>	<i>M.E. CO2Footprint</i>	5.31	(14.73)	5.35	(14.89)

Appendix 2

Appendix 2 – Parameter estimates for the measurement equations. Case study 2.

<i>Variable</i>	<i>Equation</i>	<i>LVLC</i>		<i>DCLV</i>	
<i>Scale Parameter</i>	<i>M.E. HSTrains</i>	1.13	(8.48)	1.09	(7.6)
<i>Threshold 1</i>	<i>M.E. HSTrains</i>	-3.74	(-13.3)	-3.77	(-12.42)
<i>Threshold 2</i>	<i>M.E. HSTrains</i>	-2.21	(-11.8)	-2.21	(-11.3)
<i>Threshold 3</i>	<i>M.E. HSTrains</i>	-0.179	(-1.47)	-0.16	(-1.37)
<i>Threshold 4</i>	<i>M.E. HSTrains</i>	1.43	(10.12)	1.47	(10.56)
<i>Scale Parameter</i>	<i>M.E. RailLines</i>	2.03	(3.9)	2.23	(2.84)
<i>Threshold 1</i>	<i>M.E. RailLines</i>	-2.83	(-12.73)	-2.74	(-11.66)
<i>Threshold 2</i>	<i>M.E. RailLines</i>	-1.5	(-10.01)	-1.44	(-9.44)
<i>Threshold 3</i>	<i>M.E. RailLines</i>	0.0287	(0.26)	0.0539	(0.53)
<i>Threshold 4</i>	<i>M.E. RailLines</i>	1.4	(9.91)	1.4	(9.81)