

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

Working Paper

Analyzing the continuity of attitudinal and perceptual indicators in hybrid choice models

DIW Discussion Papers, No. 1528

Provided in Cooperation with:

German Institute for Economic Research (DIW Berlin)

Suggested Citation: Bahamonde-Birke, Francisco J.; Ortúzar, Juan de Dios (2015) : Analyzing the continuity of attitudinal and perceptual indicators in hybrid choice models, DIW Discussion Papers, No. 1528, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:

<https://hdl.handle.net/10419/125152>

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.

1528

Discussion
Papers

Analyzing the Continuity of Attitudinal and Perceptual Indicators 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>

Analyzing the Continuity of Attitudinal and Perceptual Indicators in Hybrid Choice Models

Francisco J. Bahamonde-Birke^{*}

Juan de Dios Ortúzar[†]

Abstract:

This paper addresses the continuity of attitudinal and perceptual indicators in hybrid discrete choice models and the main objective of this work is to compare the consequences of treating the indicators as continuous or ordinal outcomes, given different assumptions about the way in which these are stated. Based on tradition and for computational reasons, such indicators are predominantly treated as continuous outcomes. This usually neglects their nature (as respondents are normally asked to state their preferences, or level of agreement with a set of statements, using a discrete scale) and may induce important bias.

We conducted an analysis based on simulated data and real data (two case studies) and were able to find that the distribution of the indicators (especially when associated with non-uniformly spaced thresholds) may lead to a clear deterioration of the model's predictive capacity, especially when assuming continuous indicators. Along the same line, higher relative variability among the latent variables increases the differences between both approaches (ordinal and continuous outcomes), especially concerning goodness-of-fit of the discrete-choice component. It was not possible to identify a relation between the predictive capacity of both approaches and the amount of available information.

Finally, both case studies using real data show an improvement in overall goodness-of-fit when considering the indicators as ordinal outcomes, but this does not translate in a better predictability of the discrete choices.

Keywords: Latent Variables, Hybrid Discrete Choice Modelling, Attitudinal and Perceptual Indicators, Continuity

JEL: C35, C50

^{*} Energy, Transportation and Environment Department, Deutsches Institut für Wirtschaftsforschung – Berlin and Technische Universität Berlin. E-mail: bahamondebirke@gmail.com

[†] Department of Transport Engineering and Logistics, Pontificia Universidad Católica de Chile, e-mail: jos@ing.puc.cl

1. INTRODUCTION

The study of unobserved latent variables is a hot topic nowadays. When it cannot be assumed that the outcome satisfies the property of continuity (as in the case of discrete, ordinal or nominal variables), the modeller usually relies on unobserved continuous constructs; this way, it is assumed that the underlying factors, represented through latent variables, explain the observed non-continuous output.

Discrete choice (DC) modelling (McFadden, 1974) is a special case of dealing with nominal variables. Here, it is assumed that individuals face a set of alternatives associated with a particular underlying utility function that depends on both the characteristics of the individuals and the alternatives' attributes. Observed choices are considered to be the result of a maximization process, where individuals opt for the alternative associated with the highest expected utility.

An important limitation of this framework is that it only allows for observed variables to impact utility. The effort of combining discrete choice models with other unobserved latent variables, accounting for unobserved factors that may be relevant in the decision making process, such as attitudes and perceptions, led at the 80s to propose the first versions of hybrid discrete choice (HDC) models (McFadden, 1986; Train *et al.*, 1987), an approach currently based on the Multiple Indicators Multiple Causes (MIMIC) model (Zellner, 1970; Bollen, 1989). Here it is assumed that latent variables (associated with attitudes and perceptions) explain a set of indicators previously gathered from the individuals (through so called *measurement equations*), while being explained by a set of characteristics from the users and the alternatives (through so called *structural equations*). HDC modelling basically consists in the use of latent constructs associated with a MIMIC model as explanatory variables in a DCM framework.

Despite not having a great impact in its origins (mainly due to computational issues), HDC modelling was revitalized during the last decade (Ben-Akiva *et al.*, 2002). Since then, the approach has gained popularity and has become a standard tool in travel behaviour research (v. Acker *et al.*, 2011; Ashok *et al.*, 2002; Raveau *et al.*, 2012; Alvarez-Daziano and Bolduc, 2013; Bahamonde-Birke and Ortúzar, 2014a; among many others).

Notwithstanding, the indicators are usually considered as a linear continuous expression of the latent variables (Vredin-Johansson *et al.*, 2006; Bahamonde-Birke and Ortúzar, 2014b; Yañez *et al.*, 2010; Alvarez-Daziano and Barla, 2012). This approach may induce an important bias, as respondents are normally asked to state their preferences, or level of agreement with a set of statements, using a discrete scale (Likert, 1932). Even if the modeller allows for the individuals to state continuous indicators, it is highly doubtful whether respondents would take this into consideration and decimal numbers may be underrepresented in favour of integers. Even more, it is a debatable point if the individual considers equally all values in the scale, as simplifying heuristics (Tversky and Kahneman, 1974) may cause some levels to be ignored (leading for instance to an overrepresentation of the extremes and the midpoint).

Therefore, the indicators should not be treated as discrete outcomes, but rather as ordinal ones (Daly *et al.*, 2012). However, the impact of considering the indicators as ordinal outcomes has major implications in terms of computational costs, as model estimation gets

considerable more involved, especially when considering specifications not leading to close-forms expressions for the probabilities, such as Ordered Probit model (OPM). That is the reason why researchers considering ordinal indicators, tend to rely on the Ordered Logit model (OLM; Daly *et al.*, 2012; Hess *et al.*, 2013).

2. THEORETICAL BACKGROUND

When considering a MIMIC model, the analyst assumes the existence of latent variables, which are a function of positively observed explanatory variables and, eventually, of other latent constructs (Kamargianni *et al.*, 2014; Link, 2015). This way, the above-mentioned *structural equations* may take the following form (assuming a linear specification):

$$\eta = \alpha_x \cdot X + \alpha_\eta \cdot \eta^* + \nu \quad (2.1)$$

Here, η and η^* are vectors describing sets of jointly dependent endogenous latent constructs, X a set of exogenous observed explanatory variables and ν an error term that can follow any distribution, but is usually assumed to be normally distributed with mean zero and a given covariance matrix Σ_η . α_x and α_η are matrices of parameters to be estimated.

In a MIMIC structure, this set of equations will always be unidentified (it is mandatory), so that it is necessary to consider it in conjunction with a *measurement equations* set. The latter may be described in the following manner (again assuming a linear specification):

$$I = \gamma_x \cdot X + \gamma_\eta \cdot \eta + \zeta \quad (2.2)$$

where I is a vector of exogenous indicators and ζ an error term, the distribution of which will depend on the assumptions regarding the indicators. Finally γ_x and γ_η are matrices of parameters to be estimated.

Sufficient and necessary conditions for identification of the equations' sets are well-known (Bollen, 1989) and it is necessary to constrain either some parameters or the variance associated with the error terms of the *structural equations*¹ (which is preferred in this work, given the fact that it simplifies selecting the parameters to be constrained). That being the case, the model can be estimated maximizing the likelihood function (once the structural equations have been reduced):

$$L = \int_{\eta} P(I | X, \eta; \alpha, \gamma, \zeta, \nu) \cdot f(\eta | X, \eta^*; \alpha, \nu) \cdot d\eta \quad (2.3)$$

As mentioned in the previous section, for the sake of simplicity as well as for historical reasons (the first models relying on this structure were based on that assumption; Morikawa *et al.*, 1996; Ben-Akiva *et al.*, 2002) the indicators have been considered continuously distributed. This way, assuming normally distributed error terms with mean

¹ It is important to notice that it only applies when working with more than one measurement equations. In other case it is necessary to impose one additional constrain.

zero and a diagonal covariance matrix Σ_I for the measurement equations, the probability in equation (2.3) can be depicted in the following manner:

$$P(I | X, \eta; \alpha, \gamma, \nu) = \frac{1}{\sigma_I} \phi \left(\frac{I - \gamma_X \cdot X - \gamma_\eta \cdot \eta}{\sigma_I} \right) \quad (2.4)$$

where σ_I are the elements in the diagonal of Σ_I and are parameters to be estimated. ϕ stands for the pdf of the standard normal distribution.

When the indicators are no longer considered to be a continuous response, but ordinal variables (as they indeed are), equation (2.4) does not longer apply. In this case the probability of observing a given output would be given by:

$$\begin{aligned} P(I = n | X, \eta; \alpha, \gamma, \zeta, \nu) &= P(\psi_{n-1} < \gamma_X \cdot X + \gamma_\eta \cdot \eta \leq \psi_n | X, \eta; \alpha, \gamma, \zeta, \nu) \\ &= P(\psi_n < \gamma_X \cdot X + \gamma_\eta \cdot \eta | X, \eta; \alpha, \gamma, \zeta, \nu) - P(\psi_{n-1} < \gamma_X \cdot X + \gamma_\eta \cdot \eta | X, \eta; \alpha, \gamma, \zeta, \nu) \end{aligned} \quad (2.5)$$

Here, Ψ are thresholds to be estimated and n describes a given level of the ordinal variable I . This has m different levels of I , $\Psi_0 = -\infty$ and $\Psi_m = \infty$, with the intermediate thresholds increasing monotonically. Depending on the specification of the error term ζ , which is normally assumed to be either normally or logistically distributed, with mean zero and diagonal covariance matrix Σ_ζ , equation (2.5) will lead to an OPM or OLM framework, respectively.

Combining a MIMIC model with a DC framework (considering latent variables into the utility function of one or more alternatives) leads to a HDC model, where under the assumption of additive linearity, the utility functions may be described in the following fashion:

$$U_j = \beta_x \cdot X + \beta_\eta \cdot \eta + \varepsilon \quad (2.6)$$

where β_x and β_η are vectors of parameters to be estimated and ε an error term that is usually assumed to follow an EV1 distribution with the same mean for all alternatives and a given covariance matrix. If the covariance matrix is considered to be diagonal, the choice probabilities will be given by a Logit model; if not the choice situation will be described by other member of the Generalized Logit family (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 takes a positive value. The likelihood function for the integrated framework will then take the following form:

$$L = \int_{\eta} P(I | X, \eta; \alpha, \gamma, \zeta, \nu) \cdot P(y | X, \eta; \alpha, \beta, \varepsilon, \nu) \cdot f(\eta | X, \eta^*; \alpha, \nu) \cdot d\eta \quad (2.7)$$

Observing equation (2.7), it is clear that the discrete choice component of the HDC model can be considered as just another measurement equation for the integrated framework, with the difference that its output is discrete and not ordinal. Moreover, when working with only

one indicator, it would be enough to constrain the variances of the structural equations, as the discrete choice component will provide the additional information required to identify the model as it would be another indicator.

3. SIMULATION EXERCISE

The main objective of this work is to compare the consequences of treating the indicators as continuous or ordinal outcomes, given different assumptions and distributions regarding the way in which they are stated. As both specifications cannot be directly compared statistically (given the underlying assumptions), we will conduct a qualitative analysis based on the likelihood for the overall model and for the discrete choice component only (which is the part, we aim to reproduce as accurately as possible²). Along the same line, we want to analyse if both specifications are affected by other characteristics of the sample, such as the variability of the structural equations (the importance of the stochastic part relative to the deterministic part which is independent of the normalization), the distribution of the indicators and the number of observations.

For that reason, we conducted first a simulation exercise. This allows us to analyse the aforementioned effects in a controlled environment (free of undesired effects) as well as to vary the characteristics we want to examine, when generating the samples.

3.1 Generation of the Datasets

Following Williams and Ortúzar (1982) we generated different samples of simulated individuals behaving compensatory in a binomial Logit framework, according to the following structural and utility equations:

$$\eta = \alpha_1 \cdot X_{\eta 1} + \alpha_2 \cdot X_{\eta 2} + \alpha_3 \cdot X_{\eta 3} + \alpha_4 \cdot X_{\eta 4} + \nu \quad (3.1)$$

$$\Delta U = \beta_{X1} \cdot X_1 + \beta_{\eta} \cdot \eta - \beta_{X2} \cdot X_2 + \varepsilon \quad (3.2)$$

Here, $X_{\eta 1}$ follows a discrete uniform distribution across the population, taking the values 0 and 1 (or, alternatively, Bernoulli distributed with $q = 0.5$). Similarly, $X_{\eta 2}$ follows a discrete uniform distribution in the range $[0; 2]$. $X_{\eta 3}$ and $X_{\eta 4}$ are assumed to be continuous uniformly distributed across the sample in the ranges $[0; 2]$ and $[0; 3]$, respectively. Finally, $X_{\eta 3}$ and $X_{\eta 4}$ follow normal distributions $N(4, 2)$ and $N(8, 1.5)$, respectively.

All α and β parameters were fixed to 1 (to ease the comparison of the results). The error term of the reduced utility function is independent and identically distributed following a Logistic distribution with mean 0 and scale parameter 1 (in accordance with the binomial Logit framework) for all observations. Regarding the distribution of ν we considered it to be normally distributed with mean 0 and three different cases (named Case 1, Case 2 and Case 3), which differ only on the variability of the parameter by assuming standard deviations equal to 0.5, 1 and 2, respectively.

² Focusing the analysis on parameter recovery is not expedient as magnitudes depend on the normalization, which is in turn, affected by the fact that in the estimation we are neglecting the assumptions used in the generation of the database.

We also considered two indicators the measurement equations of which are simply the sum of a latent variable η and an error term ζ , which distributes standard Normal. However, the situation is not as simple as by using a Likert-scale we only allow for the stated indicators to take integer values between 1 and 5. Therefore, it is necessary to define the way to relate this continuous equation to discrete indicators. We considered six different cases:

- a) Case A: This corresponds to the simplest assumption, i.e. normalizing the results of the measurement equations, so that all are contained between 0 and 5, and then associating the results contained in the different quintiles of the distribution to the respective level of the indicators. This case resembles the assumptions behind treating the indicators as continuous variables.
- b) Case B: Thresholds are established, so that the levels of the indicators have a uniform distribution.
- c) Case C: As above, but the distribution of the levels follows a discrete approximation of an inverted triangular distribution. This case attempts to depict an overrepresentation of the extremes.
- d) Case D: As above, but the distribution of the levels follows a discrete approximation of a triangular distribution. This case attempts to depict an overrepresentation of the intermediate levels.
- e) Case E: As above, but the distribution of the levels follows a discrete approximation of an exponential function of the form:

$$f(x) = \frac{e^x - 1}{e - 2}$$

- f) Case F: As above, but the distribution of the levels follows a discrete approximation of an exponential function of the form:

$$f(x) = e - e^x$$

Figure 1 represents graphically the different cases considered.

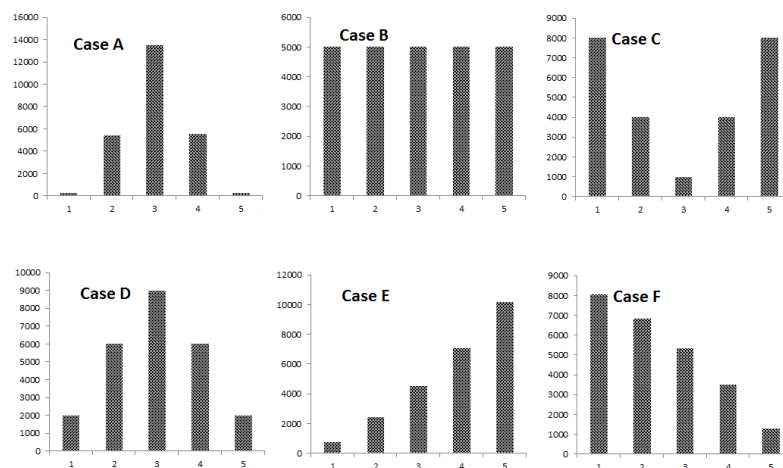


Figure 1 – Considered distributions for the indicators. Simulation exercise³.

³ The actual distribution of the indicators considered in Case A depends on the variability of the structural equations. The distribution presented in Figure 1 corresponds to Case 1A (slight variability).

Altogether we generated 18 different samples of 25,000 pseudo-individuals. Additionally, and in order to test the influence of sample size we simulated populations of 5,000; 1,000 and 500 individuals. In this case we considered all six distributions for the indicators, but only Case 2 regarding the variability of the structural equations.

3.2 Estimation results

Our first interest was the influence of the indicators' distribution as well as of the variability of structural equations on the model results. With this goal we estimated models for the 18 full-size samples, considering two different specifications: in the first one, we treated the indicators as continuous outputs and assumed that the error terms of the measurement equations distributed Normal; in the second specification, we considered the indicators to be of an ordinal nature and assumed logistically distributed error terms for the measurement equations; this leads to an OLM specification. Even though given the fashion in which the data was generated it would be more appropriate to consider an OPM specification, we opted for the logistic distribution, as this is the most common assumption when dealing with ordinal indicators in HDC models (given its computational advantages).

The models were estimated simultaneously using PythonBiogeme (Bierlaire, 2003). Taking advantage of the fact that we are dealing with only one latent variable the log-likelihood was computed via numerical integration. Table 1 presents results for the overall model for all cases analysed. The upper results correspond to the assumption of discrete indicators while the inferior ones are related to the continuous specification.

Table 1 – Goodness-of-fit of the overall model. Simulation exercise.

	<i>Case A</i>	<i>Case B</i>	<i>Case C</i>	<i>Case D</i>	<i>Case E</i>	<i>Case F</i>
<i>Case 1</i>	-44,560.6 -44,520.5	-69,244.4 -76,888.0	-63,194.6 -88,821.7	-61,367.4 -62,045.2	-58,165.0 -66,678.6	-63,240.4 -69,676.9
<i>Case 2</i>	-46,874.2 -46,886.0	-72,506.4 -79,939.7	-66,258.0 -91,551.2	-64,676.7 -65,347.4	-61,426.4 -69,365.6	-66,457.5 -72,582.8
<i>Case 3</i>	-49,234.6 -48,140.6	-74,317.8 -79,794.7	-68,402.9 -91,524.9	-66,808.3 -66,060.2	-63,671.9 -69,402.5	-68,478.3 -72,704.8

As can be observed, the log-likelihood depends strongly on the distribution of the indicators. In line with our expectations, a greater relative variability in the structural equations affects negatively the log-likelihood of the estimated model. Regarding the estimation's assumptions, the discrete specification appears to clearly outperform the assumption of continuous indicators, when the distribution of the indicators is not linearly related to the continuous results of measurement equations (Case A). Additionally this former distribution provides the best overall fit, notwithstanding the assumptions considered in the estimation. The worst goodness-of-fit is associated with Case C (overrepresentation of the extremes), which is the case departing the most from the assumptions of Case A.

To offer a better comparison between both assumptions for the indicators, Table 2 presents the results for the logarithm of direct likelihood ratio (DLR; upper cell). The table also shows the logarithm of the DLR between both approaches considering the DC component

of the model only (the likelihood of the model estimated under the OLM assumption is considered in the numerator of the DLR).

Table 2 – DLR between both approaches under different variability.

	<i>Case A</i>	<i>Case B</i>	<i>Case C</i>	<i>Case D</i>	<i>Case E</i>	<i>Case F</i>
<i>Case 1</i>	-40.13	7,644.5	2,5627.3	677.8	8,513.6	6,436.6
	-0.045	1.577	3.391	0.096	3.044	1.665
<i>Case 2</i>	11.79	7,433.3	25,293.2	670.7	7,939.2	6,125.3
	612.7	604.4	621.1	599.1	652.3	609.3
<i>Case 3</i>	-1,094.1	5,476.9	23,122.0	-748.1	5,730.6	4,226.5
	1,910.9	1,823.8	1,900.4	1,800.6	1,892.2	1,838.8

As mentioned before, the OLM specification clearly outperforms the continuous assumption in cases B, C, E and F, while the results are not conclusive for cases A and D (which present the most similar distribution of all considered cases). Nevertheless, the gap appears to diminish (or increase in favour of the continuous assumption), when the relative variability of the structural equation increases.

More telling than the overall log-likelihood, however, is to consider only the log-likelihood associated with the DC component of the model; first, this part is equal in both specifications and second, when considering a HDC, the focus remains on predicting the observed choices as accurately as possible and not the indicators. In this case, we cannot observe any significant difference when the relative variability of the structural equations is small but the gap increases dramatically as this variability gets larger (in opposition to the log-likelihood of the overall model). This finding applies to all distributions considered for the indicators.

Regarding the number of observations available for estimation, we have calibrated models considering different sample sizes, as described in the previous section. Table 3 presents the results for the logarithm of the direct likelihood ratio for the overall model and for the DC component. The results are normalized by the number of observations in order to make them comparable (the results are normalized to 500 observations).

Table 3 – Log-DLR between both approaches under different sample sizes.

<i># of Obs.</i>	<i>Case A</i>	<i>Case B</i>	<i>Case C</i>	<i>Case D</i>	<i>Case E</i>	<i>Case F</i>
<i>500</i>	5.184	164.732	513.641	14.134	184.892	142.275
	13.318	13.035	13.444	12.724	13.653	13.293
<i>1,000</i>	8.263	167.259	537.652	17.707	170.015	147.154
	17.193	17.094	17.385	16.877	16.264	17.151
<i>5,000</i>	-1.014	148.560	513.535	11.458	155.781	125.595
	10.410	10.335	10.667	10.124	14.082	10.497
<i>25,000</i>	0.236	148.666	505.863	13.414	158.785	122.506
	12.253	12.089	12.423	11.982	13.045	12.185

As can be appreciated, there is no clear tendency regarding the goodness-of-fit and the number of observations used in the estimation. Moreover the gap between both approaches

(for the overall model and for the DC component only) appears to stay constant, notwithstanding the number of observations.

Finally it must be remarked that, in this particular case, the convergence velocity of the continuous approach was about 50% superior, but it depends on the model structure and the starting values of the parameters.

4. CASE STUDY 1 – ELECTROMOBILITY IN AUSTRIA

To provide further information for our analysis, we repeated our experiment considering real data. Our first case study comes from the DEFINE project and a web-based survey conducted in Austria during February 2013. The sample is representative for the Austrian society.

The survey considered a SP-experiment and was set in the context of choosing between different options of electromobility (for details see Bahamonde-Birke and Hanappi, 2015). The vehicle purchase experiment used a labelled experimental design including four choice alternatives referring to one propulsion technology each: conventional vehicles (CV), plug-in hybrid-electric vehicles (PHEV), hybrid-electric vehicles (HEV) and electric vehicles (EV). Each of the alternatives 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 characterised 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 within 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. Figure 2 describes the distribution of the indicators and Table 4 presents an overview of the variables that are relevant to our study⁴.

As can be observed from Figure 2, the distribution of the indicators is quite variable. In some cases the mode is given by one of the central points (4), while in other the upper third is overrepresented, departing from the assumptions of the hypothesis of continuity. The shape of the distributions differs from the shapes considered in our simulation exercise, but broadly tend to depict an overrepresentation of the intermediate levels.

⁴ Bahamonde-Birke and Hanappi (2015) considered further variables and also estimated more involved models than the one considered in this study. For the purposes of this work, this specification is considered appropriate, as additional complexity would only add noise to our analysis as well as computational complexity. Bahamonde-Birke and Hanappi (2015) considered the attitudinal indicators to be a continuous expression of the latent variable.

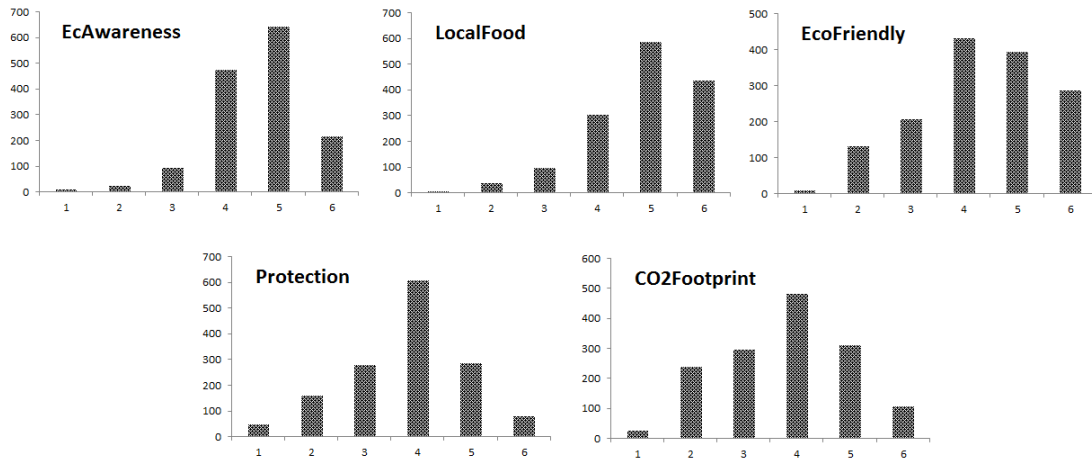


Figure 1 – Considered distributions for the indicators. Case Study 1.

Table 4 – 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.
<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. Again, taking advantage of having just one latent variable, the log-likelihood was computed using numerical integration. The results for the models assuming continuous (MCT1) and ordinal indicators (OLM1) are presented in Table 5. 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 DC component is also reported (the results for the measurement equations are omitted, as they are not suitable for comparison).

Table 5 – Parameter estimates. Case study 1.

<i>Variable</i>	<i>Equation</i>	<i>MCTI</i> ⁵	<i>OLMI</i>
<i>Vienna</i>	<i>S.E. LV Green</i>	-0.135 (-2.01)	0.138 (2.04)
<i>Male</i>	<i>S.E. LV Green</i>	-0.286 (-4.72)	0.302 (5.01)
<i>HighSkill</i>	<i>S.E. LV Green</i>	0.599 (6.64)	-0.591 (-6.58)
<i>MidSkill</i>	<i>S.E. LV Green</i>	0.346 (4.72)	-0.342 (-4.68)
<i>Old</i>	<i>S.E. LV Green</i>	0.621 (7.24)	-0.608 (-7.13)
<i>MidAge</i>	<i>S.E. LV Green</i>	0.384 (5.27)	-0.374 (-5.17)
<i>Carsharing</i>	<i>S.E. LV Green</i>	0.634 (4.59)	-0.645 (-4.62)
<i>CarUser</i>	<i>S.E. LV Green</i>	-0.346 (-6.67)	0.356 (6.85)
<i>ASC_CV</i>	<i>Utility CV</i>	0 (fixed)	0 (fixed)
<i>ASC_HEV</i>	<i>Utility HEV</i>	-0.079 (-0.38)	-0.0744 (-0.36)
<i>ASC_PHEV</i>	<i>Utility PHEV</i>	-0.456 (-2.09)	-0.454 (-2.09)
<i>ASC_EV</i>	<i>Utility EV</i>	-0.954 (-3.21)	-0.928 (-3.16)
<i>PP</i>	<i>Utility CV</i>	-1.14 (-9.4)	-1.13 (-9.42)
<i>PP</i>	<i>Utility HEV</i>	-1.71 (-24.11)	-1.71 (-24.13)
<i>PP</i>	<i>Utility PHEV</i>	-1.75 (-20.82)	-1.75 (-20.84)
<i>PP</i>	<i>Utility EV</i>	-1.29 (-12.7)	-1.28 (-12.77)
<i>MC</i>	<i>Utility CV, HEV, PHEV, EV</i>	-17.5 (-9.22)	-17.5 (-9.22)
<i>FC</i>	<i>Utility CV, HEV, PHEV, EV</i>	-18.9 (-16.37)	-18.8 (-16.39)
<i>PS</i>	<i>Utility CV</i>	0.0284 (5.76)	0.0284 (5.77)
<i>PS</i>	<i>Utility HEV</i>	0.0338 (8.32)	0.0338 (8.34)
<i>PS</i>	<i>Utility PHEV</i>	0.0373 (8.51)	0.0373 (8.54)
<i>PS</i>	<i>Utility EV</i>	0.00269 (0.71)	0.00268 (0.71)
<i>PS * Male</i>	<i>Utility CV</i>	-0.0163 (-4.21)	-0.0163 (-4.21)
<i>PS * Male</i>	<i>Utility HEV</i>	-0.0144 (-3.4)	-0.0143 (-3.38)
<i>PS * Male</i>	<i>Utility PHEV</i>	-0.0136 (-3.17)	-0.0135 (-3.16)
<i>PS * Male</i>	<i>Utility EV</i>	-0.00605 (-1.36)	-0.006 (-1.36)
<i>MidAge</i>	<i>Utility HEV</i>	-0.266 (-2.57)	-0.256 (-2.5)
<i>MidAge</i>	<i>Utility PHEV</i>	-0.389 (-3.72)	-0.379 (-3.66)
<i>MidAge</i>	<i>Utility EV</i>	-0.652 (-4.71)	-0.632 (-4.68)
<i>Old</i>	<i>Utility HEV</i>	-0.997 (-7.06)	-0.982 (-7.06)
<i>Old</i>	<i>Utility PHEV</i>	-1.26 (-8.48)	-1.25 (-8.5)
<i>Old</i>	<i>Utility EV</i>	-1.86 (-9.17)	-1.82 (-9.28)
<i>LV Green</i>	<i>Utility HEV</i>	0.591 (5.3)	-0.58 (-5.29)
<i>LV Green</i>	<i>Utility PHEV</i>	0.558 (4.88)	-0.546 (-4.87)
<i>LV Green</i>	<i>Utility EV</i>	1.03 (6.23)	-0.991 (-6.36)
<i>RA</i>	<i>Utility EV</i>	0.00326 (8.14)	0.00325 (8.18)
<i>LSMid</i>	<i>Utility EV</i>	0.163 (1.25)	0.165 (1.27)
<i>LSHigh</i>	<i>Utility EV</i>	0.694 (5.78)	0.693 (5.8)
<i>IM3</i>	<i>Utility EV</i>	0.233 (2.25)	0.232 (2.25)
<i>Log-likelihood</i>			
<i>Overall Model</i>		-16,620.586	-16,202.677
<i>Log-likelihood</i>			
<i>DC Component</i>		-6,625.468	-6,626.086

As can be observed, the differences between the parameters obtained following both approaches are minor and not statistically different. Nevertheless and in line with our expectations, the most notable differences are related to the structural equations and the parameters associated with the latent variable in the DC model. The overall goodness-of-fit of the model considering ordinal indicators is much higher than the adjustment of the model neglecting the discrete nature of the indicators. Nevertheless, it does not translate into a better goodness-of-fit for the DC-component. Even more, the model considering continuous indicators predicts the choices stated by the individuals slightly better, which is counterintuitive as this model neglects the nature of the indicators. The difference between both models is however rather small and it cannot be concluded that one model would predict choices consistently better than the other. The situation resembles the cases with low relative variability of the latent variables in the simulation exercise.

⁵ Even though we considered the same specification reported by Bahamonde-Birke and Hanappi (2015), the results may exhibit minor variations given the estimation technique (numerical integration as opposed to simulated maximum likelihood).

5. 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 see Bahamonde-Birke *et al.*, 2015). 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), 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. As in the previous case, Figure 3 and Table 6 present the distribution of the indicators and an overview of the variables considered, respectively.

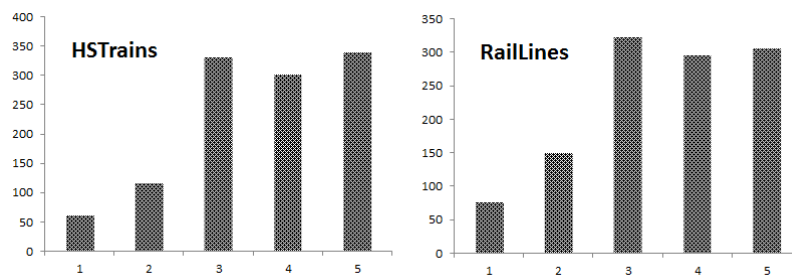


Figure 3 – Distributions for the indicators. Case Study 2.

As in Case study 1, the distribution of the indicators exhibits an overrepresentation of the upper values, but in this case both distributions are rather similar and the upper levels appear to be uniformly distributed. They can be considered to be a mixture between cases B and D of the simulation exercise.

Table 7 presents the results for models assuming continuous (MCT2) and ordinal indicators (OLM2). Again the log-likelihood was computed using numerical integration. The structure of the table is the same as in previous section.

Table 6 – 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”.

Along the line of the previous case study, the differences among the parameters calibrated following the two different specifications are not statistically significant and the major disparities are related to the structural equations models. Again, the overall goodness-of-fit of the model considering the discontinuity of the indicators is superior, but this time, in line with our expectations, the simulation exercise and the results by Daly *et al.* (2012), the proper treatment of the attitudinal indicators is also associated with a better adjustment in the DC-component. However, the difference between the goodness-of-fit for the DC-component is rather small and no conclusive result regarding the predictability of the HDC model, can be derived from the experiment.

Table 7 – Parameter estimates. Case study 2.

<i>Variable</i>	<i>Equation</i>	<i>MCT2</i>	<i>OLM2</i>
<i>Inertia</i>	<i>Utility Alternative 1</i>	0.337 (13.48)	0.337 (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>	-1.41 (-12.65)	-1.41 (-12.62)
<i>RE</i>	<i>Utility Alternative 1 and 2</i>	-0.415 (-9.59)	-0.413 (-9.52)
<i>Travel Time</i>	<i>Utility Alternative 1 and 2</i>	-0.0149 (-26.57)	-0.0149 (-26.57)
<i>Ln(Price) * Very Low Income</i>	<i>Utility Alternative 1 and 2</i>	-5.27 (-31.91)	-5.27 (-31.9)
<i>Ln(Price) * Low Income</i>	<i>Utility Alternative 1 and 2</i>	-4.69 (-26.28)	-4.69 (-26.26)
<i>Ln(Price) * Middle Income</i>	<i>Utility Alternative 1 and 2</i>	-3.62 (-14.98)	-3.62 (-15)
<i>Ln(Price) * High Income</i>	<i>Utility Alternative 1 and 2</i>	-2.52 (-7.81)	-2.54 (-7.84)
<i>Safety Level</i>	<i>Utility Alternative 1 and 2</i>	-0.00374 (-4.59)	-0.00375 (-4.59)
<i>Transfers</i>	<i>Utility Alternative 1 and 2</i>	-0.443 (-18.25)	-0.443 (-18.25)
<i>FVZ * LV TrainFan</i>	<i>Utility Alternative 1 and 2</i>	0.293 (4.81)	0.298 (4.86)
<i>Log-likelihood</i>			
<i>Overall Model</i>		-10,802.674	-10,548.544
<i>Log-likelihood</i>			
<i>DC Component</i>		-7,460.603	-7,460.161

6. CONCLUSIONS

Despite the fact that attitudinal and perception indicators normally exhibit a discrete nature (even if the modeller would allow for stating continuous values), based on tradition and on

computational reasons they are still predominantly treated as continuous outcomes. Even though the number of studies treating indicators as they are (i.e. discrete) has risen in the last years, the importance of the bias imposed by treating them continuously has not been yet extensively analysed.

We conducted a study based on simulated and real data, to examine the effects of the usual assumptions concerning attitudinal and perception indicators. Along this line, we examined the effect of the relative variability of the latent variables, the distribution of the indicators and the number of observations.

Based on simulated data we were able to show that discrete distributions (for indicators) associated with non-uniformly spaced thresholds are associated with a clear deterioration of the overall goodness-of-fit, and that this phenomenon considerably increases when the estimators are treated as continuous outcomes. A higher relative variability of the latent variables appears, however, to reduce the gap between both approaches.

When focusing exclusively on the predictive capacity of the DC-component (which is the part, modellers usually centre their efforts on), we discovered a clear worsening of the adjustment when treating the indicators continuously, as the relative variability of the latent variables increased. This deterioration applies to all distributions considered and is negligible when this variability is small. Concerning the number of observations, it was not possible to observe a clear tendency regarding the differences between both approaches and the amount of available information.

Finally, we tested these findings with real data. In both case studies analysed, the distribution of the indicators departed from the traditional assumptions underpinning the hypothesis of continuity and, as expected, treating the indicators as ordinal outcomes offers a considerably better goodness-of-fit for the overall model. Unexpectedly, this improvement does not translate into a better predictive capacity of the discrete choices, as in the case associated with small relative variability of the latent variables in the simulation exercise.

Although we were not able to establish that treating the indicators as ordinal variables when dealing with HDC models improves the model's predictive capacity (when dealing with real data), it must be stressed that considering them as continuous neglects their nature and yields worse overall model goodness-of-fit. Therefore, we recommend considering indicators as ordinal outcomes. However, this treatment is associated with higher computational costs and, given the fact that there is no evidence of significant deterioration in forecasting abilities, a trade-off may be considered.

ACKNOWLEDGMENTS

This paper is partially based on the scientific work done in the DEFINE project (Development of an Evaluation Framework for the Introduction of Electromobility). 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 Millennium 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)..

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.
- Alvarez-Daziano, R. 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, 21-22, June, 2012.
- Alvarez-Daziano, R. and Bolduc, D. (2013). Incorporating pro-environmental preferences towards green automobile technologies through a Bayesian hybrid choice model. *Transportmetrica* **9A**, 74-106.
- 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, DIW-Berlin* **1472**.
- 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, DIW-Berlin* **1474**.
- Bahamonde-Birke, F.J. and Ortúzar, J. de D. (2014a). On the variability of hybrid discrete choice models. *Transportmetrica A: Transport Science* **10(1)**, 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, Transportation Research Record: Journal of the Transportation Research Board* **2429**, 51-58.
- 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, *Proceedings of the 3rd Swiss Transportation Research Conference*, Ascona, Switzerland.
- Bollen, K.A. (1989). *Structural Equations with Latent Variables*. John Wiley and Sons, Chichester.
- Daly, A., Hess, S., Patrui, 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(2)**, 267-297.
- 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 Part D: Transport and Environment* **25**, 42-48.

- 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*.
- Link, H. (2015). Is car drivers' response to congestion charging schemes based on the correct perception of price signals?. *Transportation Research Part A: Policy and Practice* **71**, 96-109.
- McFadden, D. (1974). Conditional logit analysis of qualitative choice behaviour. In Zarembka, P. (ed.), *Frontiers in Econometrics*, 105-142. Academic Press, New York.
- McFadden, D. (1986). The choice theory approach to market research. *Marketing Science* **5(4)**, 275-297.
- Morikawa, T., Ben-Akiva, M. and McFadden, D. (1996). Incorporating Psychometric Data in Econometric Choice Models. *Working paper, Massachusetts Institute of Technology*.
- Raveau, S., Yáñez, M.F. and Ortúzar, J. de D. (2012). Practical and empirical identifiability of hybrid discrete choice models. *Transportation Research* **46B**, 1374-1383.
- Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science* **185**, 1124-1131.
- 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.
- 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. L., Ben-Akiva, M., and Bolduc, D. (2007). Identification of parameters in normal error component logit-mixture (NECLM) models. *Journal of Applied Econometrics* **22(6)**, 1095-1125.
- Williams, H.C.W.L. and Ortúzar, J. de D. (1982). Behavioural theories of dispersion and the mis-specification of travel demand models. *Transportation Research* **16B**, 167-219.
- Zellner A. (1970). Estimation of regression relationships containing unobservable variables. *International Economic Review* **11**, 441-454.