Università degli Studi di Cagliari

DOTTORATO DI RICERCA

in Economia

Ciclo XXVI

# ESSAYS ON CHOICE MODELING

Settore/i scientifico disciplinari di afferenza

SECS-P/06

Presentata da: Davide Contu

Coordinatore Dottorato Prof. Romano Piras

Tutor/Relatore Prof.ssa Elisabetta Strazzera

Esame finale anno accademico 2012 – 2013

UNIONE EUROPEA
Fondo sociale europeo

REGIONE AUTONOMA DELLA SARDEGNA

FSE 2007-2013
POR SARDEGNA
obiettivo competitività regionale e occupazione

La presente tesi è stata prodotta durante la frequenza del corso di dottorato in ECONOMIA dell'Università degli Studi di Cagliari, a.a. 2010/2013 – XXVI ciclo, con il supporto di una borsa di studio finanziata con le risorse del P.O.R. SARDEGNA F.S.E. 2007-2013 - Obiettivo competitività regionale e occupazione, Asse IV Capitale umano, Linea di Attività l.3.1 "Finanziamento di corsi di dottorato finalizzati alla formazione di capitale umano altamente specializzato, in particolare per i settori dell'ICT, delle nanotecnologie e delle biotecnologie, dell'energia e dello sviluppo sostenibile, dell'agroalimentare e dei materiali tradizionali".

Davide Contu gratefully acknowledges Sardinia Regional Government for the financial support of his PhD scholarship (P.O.R. Sardegna F.S.E. Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2007-2013 - Axis IV Human Resources, Objective l.3, Line of Activity l.3.1.).

# INDEX

**Chapter 3: Task complexity in choice experiments: A review………….................................p.107**

# Synopsis

Eliciting preferences for goods not exchanged in a market or yet to be exchanged. Obtaining the value of each of the relevant characteristics the good is decomposed into. This is what the researcher can achieve by means of choice experiments. These consist in presenting a series of choice tasks to the respondents, who are in turn asked to choose their preferred option.

Once decided to carry out a choice experiment, the researcher has to invest a lot of time and effort in two main aspects: the experimental design and model estimation. With respect to the latter, an increasing number of econometric models is available for practitioners, who have to take the responsibility of selecting the best model for the data at hand. In chapter 1 we present the results of a Monte Carlo analysis through which the performances of a series of tests for nested and non-nested models are compared, so to provide the researchers with more guidance on model selection.

The choice modeling literature has widely acknowledged the respondents might use different decision processing strategies whilst making their choice. This has a huge impact on the way we formulate the econometric models. Building on the previous literature on attribute non-attendance, in chapter 2 we focus on the possibility that what a respondent has indicated as ignored might have been instead simply less important, making use of stated attribute importance's statements. Eventually, in the empirical application presented, the information regarding the most important attribute allows to reach the best model in terms of goodness of fit and sample prediction.

Whilst setting up the experimental design the researcher has to give reasons to explain why a given number of attributes, levels, alternatives and choice tasks have been chosen, besides the experimental design's strategy itself. Designing a choice experiment implies deciding how complex each choice task is going to be. Many studies have investigated how error variance and monetary valuations are affected by varying the levels of task complexity. However, there is not a common definition of what task complexity is. Hence, in chapter 3, a definition is provided, along with a systematic literature review on the issue, calling for a greater deal of task complexity's awareness.

# **Chapter 1**: A Monte Carlo analysis of the performance of selection criteria and tests for Choice Experiments models

# A Monte Carlo analysis of the performance of selection criteria and tests for Choice Experiments models[*]

## Abstract

The increased interest in modeling preference heterogeneity in Choice Experiments (CE) data has promoted the use of choice models within the family of Logit Mixture models. However, researchers have scarce guidance on how to select the appropriate model. A Monte Carlo study is set up to analyze the performance of different information criteria and tests used to discriminate between models, either nested or non-nested. In particular, within non-nested models, four criteria and three tests are assessed, all of them based on the Kullback-Leibler Information Criterion (KLIC): the AIC, AIC3, CAIC and BIC information criteria, and the tests for non-nested models proposed by Vuong, Horowitz and Ben-Akiva and Swait. Our results indicate that some criteria (especially CAIC) work better than others; and that, when feasible, information criteria should be complemented by the Vuong test, which has a low power, but it virtually never selects the wrong model, while both the Horowitz and the Ben-Akiva and Swait tests too often provide wrong indications. The paper concludes with a CE application dealing with public acceptance of wind farms, where the indications drawn from the Monte Carlo analysis are used to inform model selection.

**Keywords**: Choice Experiments Models · Information Criteria · Model Selection · Monte Carlo Analysis · Tests

[*] This work is the result of the collaboration with Elisabetta Strazzera (University of Cagliari) and Silvia Ferrini (University of Siena). Preliminary results have been presented at the International Choice Modeling Conference, Sydney 2013 (Strazzera et al. 2013).

## Abbreviations

| | |
|---|---|
| AIC | Akaike Information Criterion |
| AIC3 | Akaike Information Criterion with 3 as penalizing factor |
| BAS | Ben-Akiva and Swait |
| BIC | Bayesian Information Criterion |
| CAIC | Consistent Akaike Information Criterion |
| CE | Choice Experiments |
| CL | Conditional Logit |
| CV | Contingent Valuation |
| DGP | Data Generating Process |
| HOR | Horowitz |
| LC | Latent Class |
| LR | Log-likelihood Ratio |
| MSC | Model Selection Criteria |
| RPL | Random Parameters Logit |
| RPL_LN | Random Parameters Logit, 5 parameters log-normally distributed, one fixed |
| RPL_N | Random Parameters Logit, 5 parameters normally distributed, one fixed |
| RPL_M | Random Parameters Logit, 3 parameters normally distributed, 2 log-normally distributed, one fixed |
| RUM | Random Utility Model |
| KLIC | Kullback-Leibler Information Criterion |

## 1. Introduction

Choice Experiment models have proved to be a powerful tool to assess market and non-market values in marketing research and in many fields of applied economics: transport, health, tourism and environmental economics. The microeconomic foundation for these models is provided by the Random Utility Model (RUM), which gives rise to different econometric specifications according to the assumptions made by the researcher for the random distribution of the utility function. The Conditional Logit (CL) model (McFadden 1974), also referred to as the Multinomial Logit (MNL) model[1], is typically the baseline specification employed in applications, but it relies on an assumption of homogeneity of preferences which can often be seen as too stringent. To some extent, this assumption can be relaxed still using a CL specification, through an "observed" heterogeneity approach, i.e. by interacting individual characteristics (socioeconomic, demographic, psychometric covariates) with the attributes of the choice alternatives; however, this approach may be unsatisfactory, since it often leaves a large amount of heterogeneity unexplained. More complex models in the broad class of Logit Mixture models have been proposed in order to improve the fit of the data and the understanding of preference heterogeneity: among them, the Random Parameters Logit (RPL), also called Mixed Logit model (Revelt and Train 1998; Train 1998; McFadden and Train 2000; Train 2003), and the Latent Class (LC) model (Swait 1994; Bhat 1997) have gained wide popularity within the community of CE practitioners.

Failure to properly account for heterogeneity in preferences may have a consequence on the welfare estimates. A meta-analysis conducted by Martinez-Cruz (2013) over 20 empirical studies, indicates that the CL specification consistently yields lower estimates than models allowing for heterogeneity. The author confirms this result through a Monte Carlo analysis, showing that the CL produces downward biased estimates, although relatively more efficient than the unbiased estimates obtained from correctly specified heterogeneous models. This somehow contrasts with Provencher and Bishop (2004), who find that the CL model does better in out-of-sample forecasts than models designed to capture preference heterogeneity. The issue surely needs further examination, but besides the problem of the validity of welfare estimates there are other reasons for being interested in a correct specification of the heterogeneity of preferences, and to understand how preferences are distributed across the population: for example, this may help a finer tuning of marketing strategies, or a better comprehension of different standpoints across the population interested in a public project.

Since the "true" data generating process is obviously unknown to the researcher, a viable strategy to find the correct (or at least the most satisfactory) model for the choice data at hand is to perform a specification search. Greene and Hensher (2003, p. 698) support this view: *"We encourage a greater effort to compare and contrast such advanced models as one approach to searching for rules on stability in explanation and prediction".* But after estimating and comparing different models, the problem is how to select the "best" specification. The tools available to the researcher for model selection vary according to the statistical structure of the models to be compared. In particular, a distinction must be made for cases in which: a) the CE models to be compared are "regularly" nested, i.e. one model can be obtained from the other after restrictions in the interior of the parameter set; b) the models are nested but the restriction parameters lie in the boundary of the parameter space; c) the models are strictly non nested, i.e. no model can be obtained from the other after restrictions in the parameter set.

When the models are "regularly" nested the testing procedure is straightforward, since classical testing procedures can be adopted. For example, in the case of CL and RPL models, a Likelihood Ratio (LR) test can be used to assess whether the RPL is a significant improvement over the restricted CL model, since the CL can be obtained from a given specification of a RPL by appropriate restrictions. This test can be applied in addition to the t-test on the significance of the standard deviations parameters, to test the joint significance of the estimated standard deviation coefficients. Alternatively, Wald or Lagrange Multiplier (LM) tests can also be applied, as discussed in Section2 below.

If the comparison is between models for which the conventional testing procedures cannot be applied, model selection is more problematic. In particular, when the models are nested, but the parameters of the restrictions to be tested are on the boundary of the parameter space, the standard regularity conditions for classical testing procedures do not hold. In the context of CE models, this is typically the case of comparison between LC models with different numbers of segments (which include the CL model as a degenerate mixture model with one segment only). For example, a two-class LC model is obtained by constraining one latent class probability in the three-class LC model to zero, i.e. at the boundary of the probability parameter space. In this case, classical statistical tests for comparison of models are not available, since the regularity conditions for the tests to be chi-squared distributed are not satisfied. A testing procedure has been proposed by Lo et al. (2001), based upon the test for overlapping models constructed by Vuong (1989) to confront adjacent latent class models (i.e. models with K and K+1 segments). However, Jeffries (2003) shows that the Lo-Mendell-Rubin test cannot be considered a valid instrument to detect the correct number of segments because the assumptions made by Vuong for comparison of overlapping models do not hold in this case; the author supplements his theoretical analysis with a Monte Carlo study, which supports the result.

Empirical applications of LC models, commented in the next section, generally use model selection criteria (MSC) based on the Kullback-Leibler Information Criterion (KLIC) to decide on the number of segments. The information criteria denote a penalized likelihood function, i.e. the estimated likelihood with a penalty term, which is a function of the number of parameters and/or the number of observations. The criteria most commonly applied in this context are: AIC (Akaike 1973), BIC (Schwartz 1978), CAIC (Bozdogan 1987), and AIC3 (Bozdogan 1994), which are characterized by different penalty terms. The criteria are described in Section 3 of this paper. These MSC can be, and in fact are, used for comparison also between strictly non nested models, such as RPL and (non degenerate) LC models, or between RPL models with different parameters distributions.

Several problems can be associated with the use of MSC. One issue is that different criteria often diverge in their indications, so that the applied researcher is still left without guidance on the decision. Thus, it would be useful to know whether some criterion is more reliable than others, so that in case of divergence the researcher can make a safer decision. A second problem of the MSC is that, as noted by Vuong (1989), they select one model deterministically, so, when feasible, a testing approach could be preferred since it recognizes the probabilistic nature of the data. The point has also been raised by Spanos (2010, p.205):

> *"…the Akaike-type model selection procedures invariably give rise to unreliable inferences because: a) they ignore the preliminary step of validating the prespecified family of models, and b) their selection amounts to testing comparisons among the models within the prespecified family but without 'controlling' the relevant error probabilities".*

While point a) raised by Spanos may hold even when a testing approach is adopted (and, as recalled by Ortùzar et al. 2012, the issue of model validation should deserve more attention in CE estimation practice, point b) is addressed by using a testing approach. Louviere et al. (2000, p. 275) put forward the need of further research on testing between non nested models for choice data and recently Hensher et al. (2012, p.365) re-confirm the importance of this research.

In this paper, we contribute to the literature on model selection methods for CE models, and we examine the reliability of available tests for discriminating between CE models, either nested or non-nested.

A Monte Carlo analysis is carried out to evaluate the performance of different tests and criteria for model selection focusing on common models used in CE applications: CL, RPL and LC specifications. The paper aims at verifying:

- whether, in model selection, some information criterion may be preferred to others;
- whether some test performs better than others in selecting between non nested models,

- whether it is useful to adopt a testing approach to confront non nested models.

Specifically, we analyze the performance of the AIC, AIC3, CAIC and BIC model selection criteria; and three tests, based on the KLIC, for strictly non nested models: the Vuong (1989), the Horowitz (1983) and the Ben-Akiva and Swait (1986). Furthermore, we will analyze the performance of the LR test in discriminating between the CL and RPL nested models.

The paper is outlined as follows. The second section briefly reviews the instruments used in the literature for model selection between CE models. Section three describes the Monte Carlo design, the models and selection criteria and tests analyzed in the study; section four presents the results; section five illustrates how these findings can be used to guide model selection in a typical case study for environmental decision making. Conclusions are provided in the last section.

## 2. Model selection in CE studies

The issue of model selection has received scarce attention in CE modelling. With the only exception of the study by Mariel et al. (2013) which will be discussed below, we are not aware of methodological studies specifically designed to study the performance of model selection methods in this field. Therefore, we will examine results obtained for classes of models which are to some extent related to those employed in CE studies.

For instance, the issue of the detection of the correct number of segments in mixture models has been explored in marketing studies. Tuma and Decker (2013) review several simulation studies that analyze the performance of alternative model selection criteria on Finite Mixture Models (FMM), a broad class of models that includes very different mixture specifications, ranging from partial least squares to multilevel latent class models, and find that in most cases the “best” criterion for determining the number of segments is AIC3. Among them, the study by Andrews and Currim (2003) deals with a finite mixture Logit model that is close enough to the structure of the LC models employed in CE applications. Their Monte Carlo analysis is designed to test the performance of 7 model selection criteria (among them AIC, AIC3, CAIC, BIC) in retaining the correct number of segments in a mixture model for multinomial choice data. They find that AIC often either underestimates or overestimates the number of segments, and that both CAIC and BIC perform worse than AIC3; though, in contrast with AIC3, CAIC and BIC never overestimate the number of segments (holding the principle of parsimony, overestimation is deemed worse than underestimation). This result is confirmed for finite mixture normal models by Brochado and Martins (2006): while AIC and AIC3 often overestimate the number of segments, the number of overestimations by BIC and CAIC is negligible. In addition, Andrews and Currim (2003) find that the performance of CAIC and BIC considerably improves when the number of observations increases, in accordance with their characteristic of consistency. This feature is particularly important, since most CE studies, especially in environmental literature, as shown in Table A in Appendix, are characterized by larger data sets than those employed in the simulation studies surveyed by Tuma and Decker.

Table A summarizes recent studies published in main environmental economics journals[2] and shows that in empirical CE applications the AIC criterion is still the criterion most commonly used to determine the number of segments in LC model specifications. In several applications AIC is complemented with BIC (Kanchanaroek et al. 2013; Nguyen et al. 2013; Strazzera et al. 2012; Hidrue et al. 2011), sometimes with AIC3 (Thiene et al. 2012; Borg and Scarpa 2010), or CAIC (Thiene et al. 2012; Meyerhoff et al. 2010). Hynes et al. (2008, p.1020) refer to conclusions drawn by Andrews and Currim (2003) to point out that *“it is unrealistic to assume that one segment-retention criterion is best for mixtures of all types of distributions in all situations”*, and conclude that it is better to use different criteria to decide on the number of segments in LC models for CE data. It would be useful to shed some light on the performance of different selection criteria in the context of finite mixture Logit models typically used in CE applications, so that practitioners, knowing the relative strengths and weaknesses of each criterion, can make a more informed decision.

The comparison between RPL and CL models is quite common in the reviewed literature and is usually based on the t-statistic tests of the deviation of the random parameters. Some studies report the log-likelihood ratio (LR) test (Birol et al. 2006; Gracia et al. 2009; Kataria 2009; Abdullah and Mariel 2010; Kosenius 2010; Van Loo et al. 2011; Gelo and Koch 2012). An alternative test for nested models, rarely applied in the literature, is the Lagrange Multiplier (LM) test proposed by McFadden and Train (2000). Mariel et al. (2013) compare the performance of the LM and the t-statistic tests in selecting the correct specification for the attribute coefficients of the RUM model in a Monte Carlo study. Their results indicate that the LM test has good properties in terms of empirical size, while the t-statistic has higher empirical size, though it also shows higher power. The authors warn *"against a straightforward and regrettably widely used selection of random parameters in an RPL model based on t-statistics of their deviations without deeper study based on an alternative test procedure"* Mariel et al. (2013, p.56). It would be interesting to have also some indications on the performance of the LR test in this setting.

The comparison between LC and RPL models is usually made through MSC, but a testing approach has also been taken in some recent studies. In particular, the Ben-Akiva and Swait (1986) test has been employed in several applications (e.g. Birol et al. 2006; Colombo et al. 2009; Shen 2009; Brouwer et al. 2010; Kosenius 2010). Burton and Rigby (2009) apply the Clark (2003) test, which is a variant of the Vuong (1989) test for non-nested models[3]. The Vuong test has been applied in Contingent Valuation studies (e.g. Czajkowski and Scasny 2010; Glenk and Fischer 2010; Genius and Strazzera 2011; Halkos and Jones 2012); however we are not aware of any application in CE studies, the only exception being Czajkowski et al. (2009), who mention that this test was used along with other selection methods, although they do not give specific information on the results of the testing procedure. Given that the distributional properties of the Clark test are unknown, while the Vuong test is known to be a consistent test, we will explore how it works for CE applications, characterized by larger data sets than typical CV applications.

Another category of CE non nested models is RPL models with different distributions specified for the random parameters. Generally the Normal distribution is used when there is no a priori sign restriction (Shen 2009) and the symmetry assumption is reasonable (Hess 2010), otherwise a Log-normal distribution is the preferred candidate. Alternative specifications for the random parameter distributions are Uniform and Triangular. A simple strategy to select the distribution of a random parameter is to look at fit improvement, as in Hynes et al. (2008), but a testing approach can also be adopted: for example, Fosgerau and Bierlaire (2007) propose a semi non parametric test, but also the tests for non-nested models mentioned above could be used, hence it is useful to see how these tests work in this context.

In conclusion, a simulation study explicitly designed for CE models would give useful guidance to practitioners for selection between models commonly applied in CE studies. This is what we are describing in the following sections.

## 3. Methods

### 3.1 The Monte Carlo Design

A Monte Carlo analysis was carried out to assess the performance of different methods in discriminating between alternative estimators. We simulated individuals' choices using the frame of a CE study conducted to assess the social acceptability of wind farms in Sardinia, Italy (see section 5 in this paper for details). The study is a typical CE application for eliciting environmental preferences. Two unlabelled options are characterized by six attributes: four attributes with three levels, one with two and the monetary attribute with four levels (see Table B in Appendix for details). Each respondent values six choice sets resulting from the blocking of a $D_b$-efficient design (see Strazzera et al. 2012 for details). The study sample size was 432 individuals for 2592 choices. The Monte Carlo exercise uses the same design and sample size of the case study, and simulates the choice data considering different data generation processes (DGP); alternative estimators are then applied to the simulated data. The code for the analysis, including algorithms for estimation of the CL and LC models, has been written by the authors in GAUSS Aptech programming language, while for estimation of the RPL model we used the GAUSS routine made available by K. Train (1999)[4].

The method consists in simulating respondents' choices by first specifying the utility components, and applying different error terms based on different distributional assumptions. The application involved comparisons of two alternative scenarios, hence two utility functions are specified for each choice situation, and the simulated choice is assigned to the alternative which provides the highest level of utility, in line with the RUM theory (McFadden 1974).

Specifically, the two utilities are specified in the following way:

$$U_{i1} = \beta_{ik} X'_1 + \varepsilon_{i1} \tag{1}$$

$$U_{i2} = \beta_{ik} X'_2 + \varepsilon_{i2} \tag{2}$$

where $U_{i1}$ is the utility of individual i, $X_1$ is the $n \times k$ matrix of regressors (k attributes) and $\varepsilon_{i1}$ the random component drawn from a given distribution, attached to alternative 1; analogously $U_{i2}$, $X_2$ and $\varepsilon_{i2}$ with respect to alternative 2. Finally $\beta_{ik}$ is the $1 \times k$ vector of coefficients, which differ depending on the DGP as shown in Table 1. Hence, the choice is assigned to alternative 1 (2) if and only if $U_{i1} > U_{i2}$ ($U_{i2} > U_{i1}$). The error terms have been specified as Gumbel distributions; all simulations have been conducted using a Lower variance scenario (Gumbel standard deviation =2) and a Higher variance scenario (Gumbel standard deviation = 3).

Several DGPs are considered in turn, and the first scenario is the one with no significant heterogeneity in preferences. In particular, we set the utility parameters as fixed coefficients, and the error terms are identically and independently Gumbel distributed.

Under this setting, the probability that individual *i* chooses alternative 1 (alternative 2 otherwise), is given by:

$$P(U_{i1} > U_{i2}) = P(\beta_k X'_1 - \beta_k X'_2 + \varepsilon_{i1} - \varepsilon_{i2}) > 0 \tag{3}$$

This is the Conditional Logit Model DGP.

In order to consider situations in which heterogeneity is present, we use the Random Parameters Logit and the Latent Class models. Both estimators account for heterogeneity, since it is relaxed the CL assumption that the parameters are fixed and equal for all individuals. In a nutshell, the difference between the RPL and the LC model is that the former assumes that the parameters are random variates that follow a continuous distribution, while the latter assumes that the distribution is discrete, with same utility parameters within classes, but that differ across classes.

Considering the RPL specification first, the utility depends also on a random component included in the parameters. Hence the coefficient $\beta_{ik}$ in equations (1) and (2) is specified as follows:

$$\beta_{ik} = \overline{\beta}_k + \eta_{ik} \tag{4}$$

where $\overline{\beta}$ stands for the population mean and $\eta_{ik}$ is an error term, which is the same across choice situations for each individual. In the experiments using RPL NORMAL DGP, it is constructed as

$$\eta_{ik} = \varepsilon_{ik} \cdot \sigma_{ik} \tag{5}$$

where $\varepsilon_{ik}$ is distributed as a Normal (0,1), and $\sigma$ is a scale parameter different for each random coefficient, as reported in Table 1. For the RPL LOGNORMAL DGP the error term is constructed similarly, replacing $\eta_{ik}$ by $exp\ (\eta_{ik})$.

In our experiments we set five attributes to be randomly distributed, while the cost parameter is kept fixed. Equations (1) and (2) become:

$$U_{i1} = \sum_k (\overline{\beta}_k + \eta_{ik})\, X'_{1k} + \varepsilon_{i1} \tag{6}$$

$$U_{i2} = \sum_k (\overline{\beta}_k + \eta_{ik})\, X'_{2k} + \varepsilon_{i2} \quad . \tag{7}$$

Turning to the LC models, in this study we consider two LC DGP scenarios, namely with two and three classes, where in both cases we assign individuals to a class depending on a covariate $Z_i$ ($n \times 1$ vector). Specifically, heterogeneity is modeled as follows. First, for each individual we set:

$$I_i = \gamma_1\, Z'_i + \varepsilon_i \tag{8}$$

where $\varepsilon_i$ are errors drawn from a logistic distribution and $\gamma_1$ is a scalar. Then, we set an indicator function which groups individuals into the two (or three) classes depending on the values of $I_i$. Finally, depending on which class the individual belongs to, a set of coefficients $\beta_{k|s}$ is assigned to the attributes as well as the error terms $\varepsilon_{ij|s}$, which are drawn from a Gumbel distribution. Therefore in this setting the utility functions are specified as follows:

$$U_{i1|s} = \beta_{k|s} X'_1 + \varepsilon_{i1|s} \tag{9}$$

$$U_{i2|s} = \beta_{k|s} X'_2 + \varepsilon_{i2|s}. \tag{10}$$

Summarizing, we designed four different data generating processes (DGP): CL, RPL, and LC with two and three classes. For the RPL model, we set five attributes to be normally or log-normally distributed, while the sixth attribute, the price, was fixed. Within the LC context, we segmented respondents into groups depending on a covariate. We then estimate, for each DGP, the four models described: in this way we will always have a comparison of a model closer to the true against a misspecified model. Table 1 shows the coefficients which enter the utility function for each DGP in the study.

Table 1. Data Generating Process: Parameters and Regressors

| DATA GENERATING PROCESS | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Coefficients | CL | RPL | LC-2 classes[a] | | LC-3 classes[b] | | | Regressors | Levels |
| | | | Class1 | Class2 | Class1 | Class2 | Class3 | | |
| $\beta_1$ | 0.34 | 0.52 | -0.88 | 2.8 | 0.88 | 2.8 | 0.12 | $X_1$ | 1-2-3 |
| $\beta_2$ | 0.47 | 0.55 | -1.5 | 3.9 | 1.5 | 1.9 | 0.13 | $X_2$ | 1-2-3 |
| $\beta_3$ | 0.33 | 0.2 | 1.4 | -2.5 | 1.4 | -2.5 | 4.4 | $X_3$ | 0-1 |
| $\beta_4$ | 0.09 | 0.77 | -0.29 | 3.8 | -0.29 | 1.8 | -1.18 | $X_4$ | 1-2-3 |
| $\beta_5$ | 0.16 | 0.35 | 0.05 | 4.39 | 0.05 | 0.39 | -2.21 | $X_5$ | 0-1-2 |
| $\beta_6$ | 1.16 | 2.25 | 0.2 | 5.3 | 0.69 | 5.3 | 9.85 | $X_6$ | 0-.1-.3-.5 |
| $\sigma_1$ | | 0.68 | | | | | | | |
| $\sigma_2$ | | 0.166 | | | | | | | |
| $\sigma_3$ | | 0.51 | | | | | | | |
| $\sigma_4$ | | 0.77 | | | | | | | |
| $\sigma_5$ | | 0.35 | | | | | | | |
| $\gamma_1$ | | | 0.4 | | 0.4 | | | $Z_{LC2}$<br>$Z_{LC3}$ | 1-2-3[a]<br>1-2-3-4[b] |
| *Average class probabilities* | | | 0.501 | 0.499 | 0.36 | 0.452 | 0.188 | | |

[a]DGP LC-2 calsses. [b]DGP LC-3 classes.

## 3.2 Model Selection

In this section we briefly review some model selection criteria and tests based on the Kullback-Leibler (1951) Information Criterion[5], drawing largely from Genius and Strazzera (2001).

Given the true conditional density, unknown to the researcher, and defined as:

$$l_0(y|x) = \prod \varphi_0(y_i|x_i) \qquad (11)$$

The researcher will specify a given parametric model defined as follows:

$$l(y|x;\beta) = \prod \varphi(y_i|x_i,\beta)\,, \beta \in B \qquad (12)$$

Hence, the KLIC can be defined in the following way:

$$K_n\big(l(y|x;\ \beta\ )/l_0(y|x)\big) = {}^1\!/_n\, E_0\left(log\,\frac{l_0(y|x)}{l(y|x;\ \beta\ )}\right) \tag{13}$$

Given this measure of proximity, we can compare pairs of competing parametric models, characterized by underlying conditional densities. Furthermore, (17) can be approximated by the following measure (Gourieroux and Monfort 1995):

$$\widetilde{K} = {}^1\!/_n \sum log\varphi_0(y_i|x_i) - {}^1\!/_n \sum log\varphi\big(y_i\big|x_i;\hat{\beta}\big) \tag{14}$$

In comparing two models, for example *f* and *g*, equation (18) reduces to a log-likelihood ratio since the term ${}^1\!/_n \sum log\varphi_0(y_i|x_i)$ drops, therefore we would have:

$$LR\big(\hat{\theta},\hat{\gamma}\big) = [\sum \mathrm{log}f\big(y_i\big|x_i,\hat{\theta}\ \big) - \sum \mathrm{log}g(y_i|x_i;\hat{\gamma})] \tag{15}$$

When considering two competing nested models, respectively having $\hat{\theta}$ and $\hat{\gamma}$ as maximum likelihood estimators, the log-likelihood ratio test of the two models is given by:

$$LRT\big(\hat{\theta},\hat{\gamma}\big) = -2[\sum \mathrm{log}f\big(y_i\big|x_i,\hat{\theta}\ \big) - \sum \mathrm{log}g(y_i|x_i;\hat{\gamma})] \tag{16}$$

Under the null hypothesis, (16) follows a chi-square distribution (and some regularity assumptions, see Wooldridge (2010, p.428)) with degrees of freedom given by the difference between the number of parameters in the two models considered. In our study we employ this test to discriminate between Nested models: the RPL and the CL model (the latter nested in the former).

For comparison of “boundary” nested models (LC models with different numbers of segments) and of strictly non-nested models, we will use the following MSC: AIC (Akaike 1987), AIC3 (Bozdogan 1994), BIC (Schwartz 1978) and CAIC (Bozdogan 1987), which are defined as follows:

$$AIC = \big(2\sum logf\big(y_i\big|x_i,\hat{\theta}\ \big) - 2p\big) \tag{17}$$

$$AIC3 = \big(2\sum logf\big(y_i\big|x_i,\hat{\theta}\ \big) - 3p\big) \tag{18}$$

$$CAIC = \left(2\sum logf\left(y_i \middle| x_i, \hat{\theta}\ \right) - p(log(n) + 1)\right) \tag{19}$$

$$BIC = \left(2\sum logf\left(y_i \middle| x_i, \hat{\theta}\ \right) - p\big(log(n)\big)\right) \tag{20}$$

where *p* indicates the number of parameters and *n* the number of observations. For each pair of models under comparison, each MSC will indicate the preferred model on the basis of the highest penalized log-likelihood.

In addition, for strictly non nested models we adopt a testing approach, still based on the KLIC. In particular, the Vuong test indicates whether the two competing models are equally close to the true model, where the KLIC provides a measure of distance.

Specifically, the Vuong null hypothesis postulates that models *f* and *g* are equivalent:

$$H_0 \colon E_0\left[log\frac{f(y_i|x_i;\theta^*)}{g(y_i|x_i;\gamma^*)}\right] = 0 \tag{21}$$

whereas the alternatives are that the model *f* is better than *g* (22) or vice versa (23):

$$H_A : E_0\left[log\frac{f(y_i|x_i;\theta^*)}{g(y_i|x_i;\gamma^*)}\right] > 0, \text{ or} \tag{22}$$

$$H_A : E_0\left[log\frac{f(y_i|x_i;\theta^*)}{g(y_i|x_i;\gamma^*)}\right] < 0 \tag{23}$$

where $\theta^*$ and $\gamma^*$ stand for the pseudo-true values. Therefore this test allows the researcher to discriminate between two non-nested models in a testing framework, whose output might be: i) selecting model *f*; ii) selecting model *g*; iii) not selecting any model. Having shown the hypothesis to be tested, the Vuong test statistic is:

$$^1\!/_{\sqrt{n}} LR(\hat{\theta}, \hat{\gamma})/\hat{\omega} \tag{24}$$

where LR has been defined in (15) and $\hat{\omega}$ is:

$$\widehat{\omega} = \sqrt{\frac{1}{n}\sum\left[log\frac{f(y_i|x_i;\widehat{\theta})}{g(y_i|x_i;\widehat{\gamma})}\right]^2 - \sum\left[\frac{1}{n}log\frac{f(y_i|x_i;\widehat{\theta})}{g(y_i|x_i;\widehat{\gamma})}\right]^2} \quad (25)$$

Vuong shows that under the null hypothesis, the test in (24) converges in distribution to a standard normal. Moreover, the author indicates that the same distributional assumption holds when the LR in (24) is adjusted with a penalization terms, as those presented in (17)-(20).

Alternatively, it is possible to apply the test proposed by Horowitz (1983), where the following adjusted log-likelihood ratio index was proposed to compare two non-nested models *f* and *g*:

$$\rho_f^2 = 1 - \frac{\log f(y_i|x_i,\widehat{\theta}\,) - (K_f/2)}{\log L(y_i,0)} \quad (26)$$

Under the null hypothesis that the model with lower $\rho$ is the true model, the bound specified in (27) holds:

$$\Pr[\,\rho_g^2 - \rho_f^2 \geq z] \leq \Phi[-\sqrt{\left(-2z\log L(y_i,0)\right)}] \quad (27)$$

Finally, Ben-Akiva and Swait (1986) propose a variant of the Horowitz test. Define $\log L(y_i, 0)$ as the equal probabilities sample log-likelihood. Compute for each model, *f* and *g*, the Akaike adjusted likelihood ratio index given by:

$$\rho_f^2 = 1 - \frac{\log f(y_i|x_i,\widehat{\theta}\,) - K_f}{\log L(y_i,0)} \quad (28)$$

and analogously compute $\rho_g^2$ (K is the number of parameters). Define *z* the positive difference between $\rho_g^2$ and $\rho_f^2$; then, under the null hypothesis that *f* is the true model, the probability that the difference is greater than or equal to *z* is bounded as follows:

$$\Pr[\,\rho_g^2 - \rho_f^2 \geq z] \leq \Phi[-\sqrt{\left(-2z\log L(y_i,0) + \left(K_g - K_f\right)\right)}] \quad (29)$$

where

$$z \geq Max\ [0, \frac{K_g - K_f}{2\log L(y_i, 0)}] \tag{30}$$

and Φ is the standard normal cumulative distribution function.

In this study we analyze the performance of MSC and tests summarized in Table 2. According to the nature of the relationship between models (Nested, "Boundary" Nested or Non Nested models) different tests and MSC will be used.

**Table 2. Models relation and MSC and Tests**

| | Model Selection Criteria | | | | BAS[a] | Hor[b] | V[c] | | | | LR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | AIC | AIC3 | CAIC | BIC | | | AIC | AIC3 | CAIC | BIC | |
| Nested Models<br>(RPL vs CL) | | | | | | | | | | | ● |
| Non Nested models<br>(LC vs RPL) | ● | ● | ● | | ● | ● | ● | ● | ● | ● | |
| "Boundary" Nested models<br>(CL vs LC 2 vs LC 3 ) | ● | ● | ● | ● | | | | | | | |

Note: a) BAS is for Ben-Akiva and Swait; b) Hor for Horowitz; c) V is for Voung.

## 4. Results

The design of the Monte Carlo analysis involved 1000 replications per DGP. The results will be described by a set of Figures where different selection methods are contrasted. In some exercises (especially when the LC models are involved) the optimization algorithm failed to converge, and the corresponding replications had to be discarded: details about discharged values for model estimated for a specific DGP will be reported underneath the figures below. The assessment will always involve a pair-wise comparison between a correctly specified model and a misspecified model. The blue color in the graphs is used for correct indications given by a specific selection method; the red color stands for wrong indications; the purple color indicates cases in which the

Vuong test does not provide any indication. A summary chart of comparisons is provided in Table 3.

Table 3. Summary of comparisons

| | Data generation process | | | |
|---|---|---|---|---|
| | CL | RPL | LC-2 classes | LC-3 classes |
| Models comparison | CL-LC 2 (Figure 5) | RPL-LC 2 (Figure 3) | LC 2-CL (Figure 7) | LC 3-CL (Figure 8) |
| | CL-LC 3 (Figure 6) | RPL-LC 3 (Figure 4) | LC 2-LC3 (Figure 9) | LC 3-LC 2 (Figure 10) |
| | CL-RPL (Table4) | RPL-CL (Table 4) | LC 2-RPL (Figure 1) | LC 3-RPL (Figure 2) |

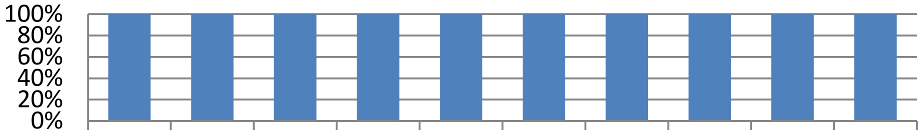
The discussion of results begins with the assessment of MSC and tests in selection of strictly non nested models characterized by different statistical structures (RPL and LC); afterwards, we will examine the performance of MSC in discriminating between "boundary" nested models, in the category of LC models (CL and LC with 2 and 3 classes); then we will compare the MSC and testing approach for selecting between RPL models with different distributions for the random parameters, to finish off with the analysis of the performance of the likelihood ratio test in selecting between the nested CL and RPL models.

Figures 1, 2, 3 and 4 report comparisons of non-nested models, RPL and LC, by means of MSC and tests. Figure 1 considers the DGP LC-2 classes and the comparison of the LC-2 classes model with the RPL. All Model Selection Criteria and tests indicate correctly the true model in both the lower and higher variance settings.

A similar positive result is obtained in the case of lower variance for the DGP LC-3 classes. However, when the higher variance scenario is considered, results are less encouraging. The CAIC and BIC MSC criteria often select the wrong model, while the Vuong test is unable to discriminate between models. On the contrary, the AIC and AIC3 criteria and the BAS and Horowitz tests perform well.

## Fig. 1. DGP Latent 2 Classes: Latent 2 Classes vs Random Parameters Logit.

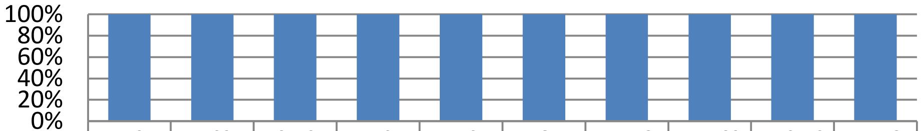
Fig. 1a. Lower variance scenario



| | AIC | AIC3 | CAIC | BIC | BAS | HOR | V_AIC | V_AIC3 | V_CAIC | V_BIC |
|---|---|---|---|---|---|---|---|---|---|---|
| 🟦 LC2 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 🟥 RPL | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 🟪 Ho not rejected | | | | | | | 0 | 0 | 0 | 0 |

*Note: N. of valid replications 746*

Fig 1b. Higher variance scenario

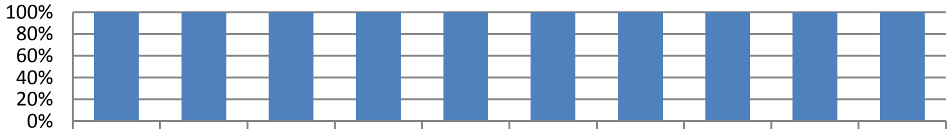


| | AIC | AIC3 | CAIC | BIC | BAS | HOR | V_AIC | V_AIC3 | V_CAIC | V_BIC |
|---|---|---|---|---|---|---|---|---|---|---|
| 🟦 LC2 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 🟥 RPL | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 🟪 Ho not rejected | | | | | | | 0 | 0 | 0 | 0 |

*Note: N. of valid replications 824*

## Fig. 2. DGP Latent 3 Classes: Latent 3 Classes vs Random Parameters Logit.

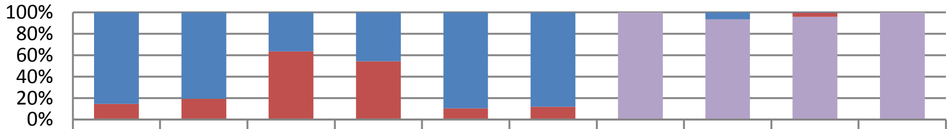
Fig. 2a. Lower variance scenario



| | AIC | AIC3 | CAIC | BIC | BAS | HOR | V_AIC | V_AIC3 | V_CAIC | V_BIC |
|---|---|---|---|---|---|---|---|---|---|---|
| 🟦 LC3 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 🟥 RPL | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 🟪 Equivalent | | | | | | | 0 | 0 | 0 | 0 |

*Note: N. of valid replications 787*

Fig. 2b. Higher variance scenario



| | AIC | AIC3 | CAIC | BIC | BAS | HOR | V_AIC | V_AIC3 | V_CAIC | V_BIC |
|---|---|---|---|---|---|---|---|---|---|---|
| 🟦 LC3 | 85.35 | 80.47 | 36.28 | 45.59 | 89.4 | 88 | 0 | 6.4 | 0.3 | 0 |
| 🟥 RPL | 14.65 | 19.53 | 63.72 | 54.41 | 10.6 | 12 | 0 | 0 | 3.7 | 0 |
| 🟪 Equivalent | | | | | | | 100 | 93.96 | 96 | 100 |

*Note: N. of valid replications 430*

It can be observed that in this case the number of discarded replications is quite high, since in many cases the estimation algorithm for the LC-3 classes model failed to converge. This leaves us with a smaller number of valid replications for this exercise, although still comparable with other Monte Carlo studies in this field[6]. It seems that the high variance of the error term induces a variability of the choice data that confounds the three classes segmentation, so that the LC-3 classes model is either unable to fit the data (it does not converge) or, if it does, the fit is not as good, and the difference between the two competing models is not so substantial, as it was in the low variance scenario.

From Figure 3 appears that the BAS and the Horowitz tests select the latent class models also when the DGP is RPL, hence giving a wrong indication. The low power of the Vuong test is seen again in these experiments, with some exceptions when the test uses the CAIC penalization. It is noteworthy that the Vuong test never selects the misspecified model, although too often it fails to reject the null hypothesis of no significant difference between the two specifications.

**Fig. 3. DGP RPL: Random Parameters Logit vs Latent 2 Classes**

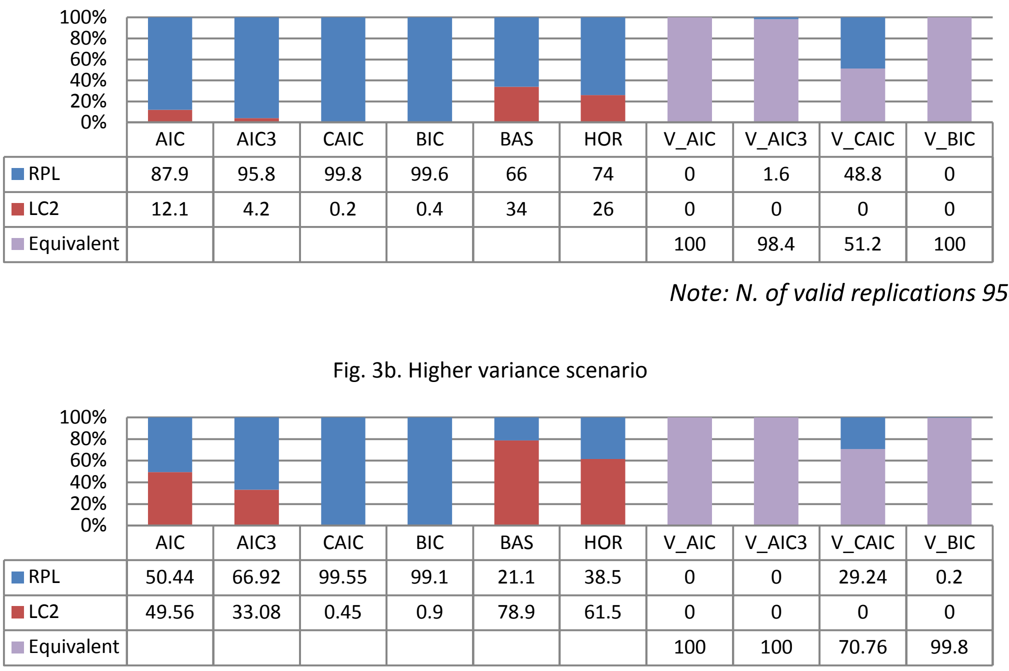


Fig. 3a. Lower variance scenario

| | AIC | AIC3 | CAIC | BIC | BAS | HOR | V_AIC | V_AIC3 | V_CAIC | V_BIC |
|---|---|---|---|---|---|---|---|---|---|---|
| 🟦 RPL | 87.9 | 95.8 | 99.8 | 99.6 | 66 | 74 | 0 | 1.6 | 48.8 | 0 |
| 🟥 LC2 | 12.1 | 4.2 | 0.2 | 0.4 | 34 | 26 | 0 | 0 | 0 | 0 |
| 🟪 Equivalent | | | | | | | 100 | 98.4 | 51.2 | 100 |

*Note: N. of valid replications 954*

Fig. 3b. Higher variance scenario

| | AIC | AIC3 | CAIC | BIC | BAS | HOR | V_AIC | V_AIC3 | V_CAIC | V_BIC |
|---|---|---|---|---|---|---|---|---|---|---|
| 🟦 RPL | 50.44 | 66.92 | 99.55 | 99.1 | 21.1 | 38.5 | 0 | 0 | 29.24 | 0.2 |
| 🟥 LC2 | 49.56 | 33.08 | 0.45 | 0.9 | 78.9 | 61.5 | 0 | 0 | 0 | 0 |
| 🟪 Equivalent | | | | | | | 100 | 100 | 70.76 | 99.8 |

*Note: N. of valid replications 892*

The MSC work well when the RPL DGP is built with lower variance, although the AIC3 and especially the AIC give often the wrong indication in the higher variance setting. The CAIC and BIC criteria seem the most efficient even with higher variance. In this comparison, the BAS test is the worst method mirrored by the Horowitz test which performs slightly better. The Vuong test penalized by CAIC is the most promising in its category.

The bad performance of the BAS and Horowitz tests is again confirmed by the results reported in Figure 4. In this case we compare the RPL and the LC model with 3 classes when the DGP is the RPL. In both the lower and the higher variance context the two tests almost always select the wrong model. The AIC is quite unsatisfactory too, while the BIC and CAIC criteria always select the correct model. It is interesting to note that in this case also the Vuong test, using the CAIC penalization works very well; the Vuong test with other corrections is less powerful, but again it never selects the wrong model.

A comparison most commonly seen in the environmental literature (see Table A in Appendix) is between the CL and LC models. These comparisons are presented in Figures 5 to 8. In this case the tests for non-nested models cannot be applied and hence only MSC are reported. Results show that when the DGP is CL, especially the CAIC and BIC criteria, but also the AIC3, work very well. However, in the comparison of CL with the LC-2 classes the AIC points in several cases to the misspecified model.

Fig. 4. DGP RPL: Random Parameters Logit vs Latent 3 Classes

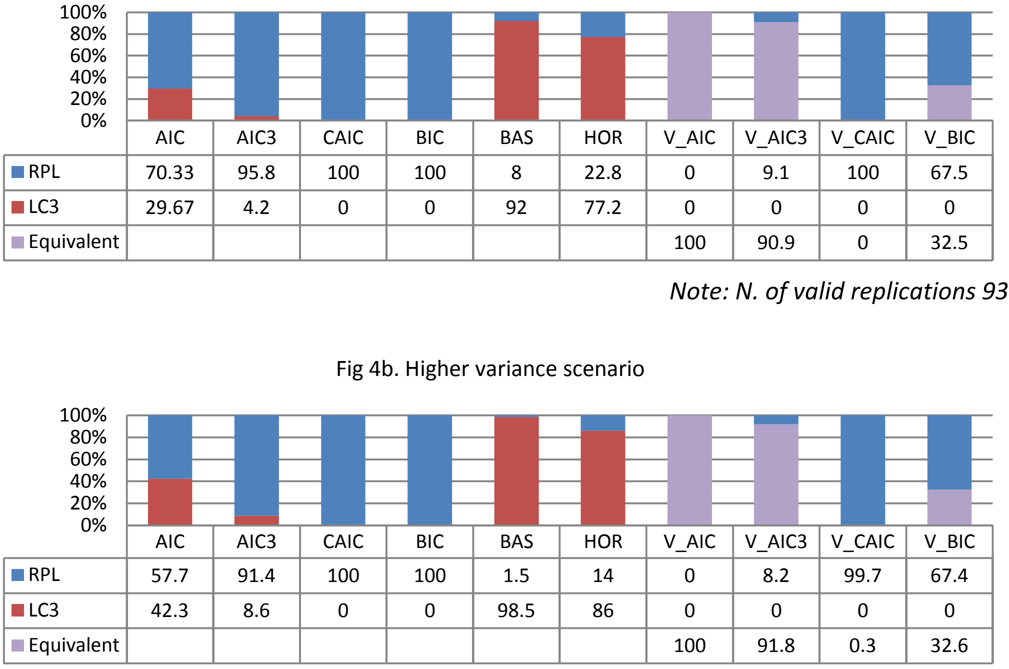


Fig 4a. Lower variance scenario

| | AIC | AIC3 | CAIC | BIC | BAS | HOR | V_AIC | V_AIC3 | V_CAIC | V_BIC |
|---|---|---|---|---|---|---|---|---|---|---|
| 🟦 RPL | 70.33 | 95.8 | 100 | 100 | 8 | 22.8 | 0 | 9.1 | 100 | 67.5 |
| 🟥 LC3 | 29.67 | 4.2 | 0 | 0 | 92 | 77.2 | 0 | 0 | 0 | 0 |
| 🟪 Equivalent | | | | | | | 100 | 90.9 | 0 | 32.5 |

*Note: N. of valid replications 934*

Fig 4b. Higher variance scenario

| | AIC | AIC3 | CAIC | BIC | BAS | HOR | V_AIC | V_AIC3 | V_CAIC | V_BIC |
|---|---|---|---|---|---|---|---|---|---|---|
| 🟦 RPL | 57.7 | 91.4 | 100 | 100 | 1.5 | 14 | 0 | 8.2 | 99.7 | 67.4 |
| 🟥 LC3 | 42.3 | 8.6 | 0 | 0 | 98.5 | 86 | 0 | 0 | 0 | 0 |
| 🟪 Equivalent | | | | | | | 100 | 91.8 | 0.3 | 32.6 |

*Note: N. of valid replications 892*

Fig. 5. DGP CL: Conditional Logit vs Latent 2 Classes

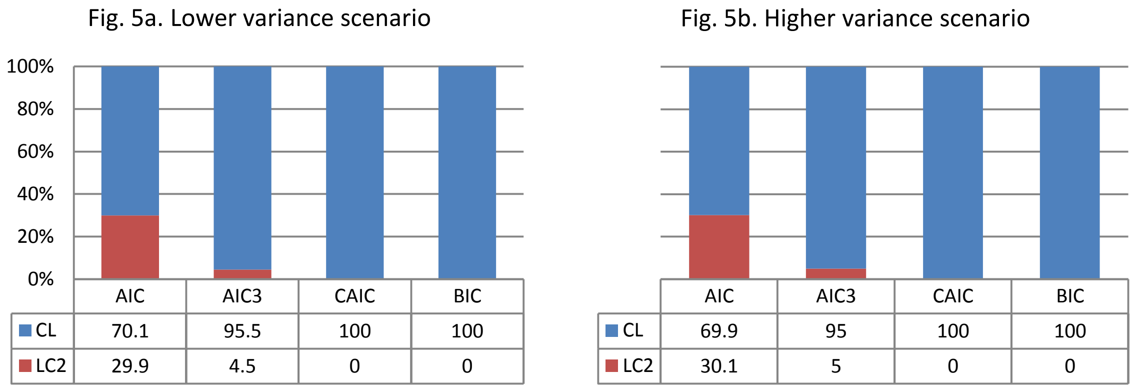


Fig. 5a. Lower variance scenario

|  | AIC | AIC3 | CAIC | BIC |
|---|---|---|---|---|
| 🟦 CL | 70.1 | 95.5 | 100 | 100 |
| 🟥 LC2 | 29.9 | 4.5 | 0 | 0 |

Fig. 5b. Higher variance scenario

|  | AIC | AIC3 | CAIC | BIC |
|---|---|---|---|---|
| 🟦 CL | 69.9 | 95 | 100 | 100 |
| 🟥 LC2 | 30.1 | 5 | 0 | 0 |

*Note: N. of valid replications 652*

*Note: N. of valid replications 671*

All criteria instead discriminate very well between the correctly specified CL model and the misspecified LC-3 classes model, both in the lower and in the higher variance scenario.

Fig. 6. DGP CL: Conditional Logit vs Latent 3 Classes

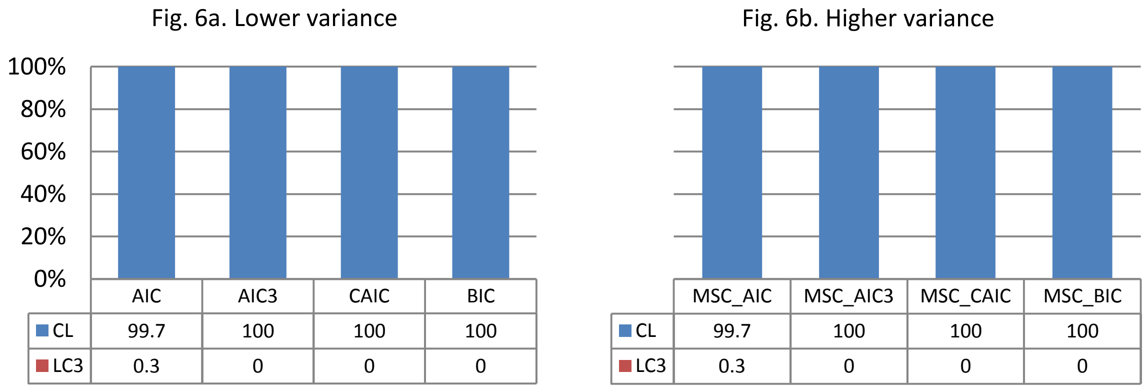


Fig. 6a. Lower variance

|  | AIC | AIC3 | CAIC | BIC |
|---|---|---|---|---|
| 🟦 CL | 99.7 | 100 | 100 | 100 |
| 🟥 LC3 | 0.3 | 0 | 0 | 0 |

Fig. 6b. Higher variance

|  | MSC_AIC | MSC_AIC3 | MSC_CAIC | MSC_BIC |
|---|---|---|---|---|
| 🟦 CL | 99.7 | 100 | 100 | 100 |
| 🟥 LC3 | 0.3 | 0 | 0 | 0 |

*Note: N. of valid replications 652*

*Note: N. of valid replications 671*

The same clear-cut situation is seen in Figure 7 where the correctly specified LC-2 classes is confronted with the CL model.

**Fig. 7. DGP Latent 2 Classes: Latent 2 Classes vs Conditional Logit**

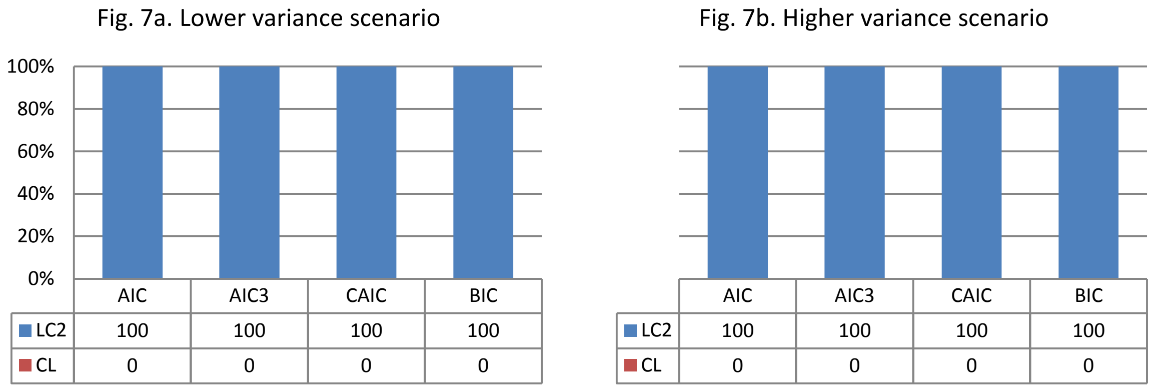


Fig. 7a. Lower variance scenario

| | AIC | AIC3 | CAIC | BIC |
|---|---|---|---|---|
| 🟦 LC2 | 100 | 100 | 100 | 100 |
| 🟥 CL | 0 | 0 | 0 | 0 |

*Note: N. of valid replications 746*

Fig. 7b. Higher variance scenario

| | AIC | AIC3 | CAIC | BIC |
|---|---|---|---|---|
| 🟦 LC2 | 100 | 100 | 100 | 100 |
| 🟥 CL | 0 | 0 | 0 | 0 |

*Note: N. of valid replications 824*

Figure 8 reports the DGP LC-3 classes: in the lower variance setting the LC model is easily selected over the CL model. Conversely, in the higher variance setting the CAIC and BIC are dominated by the other two criteria, since they choose in the vast majority of cases the CL model, i.e. they do not recognize the segmentation. This outcome is consistent with what described in Fig. 2b (DGP LC-3, comparison of LC-3 classes model with RPL): just as in that case, the LC-3 estimator does not provide a very good fit to the data, probably because the error disturbance blurs the segment separation, and the comparative advantage of the correct model in terms of KLIC is slight. In this situation the CAIC and BIC select the most parsimonious model, while AIC3 and AIC select the correct specification.

**Fig. 8. DGP Latent 3 Classes: Latent 3 Classes vs Conditional Logit**

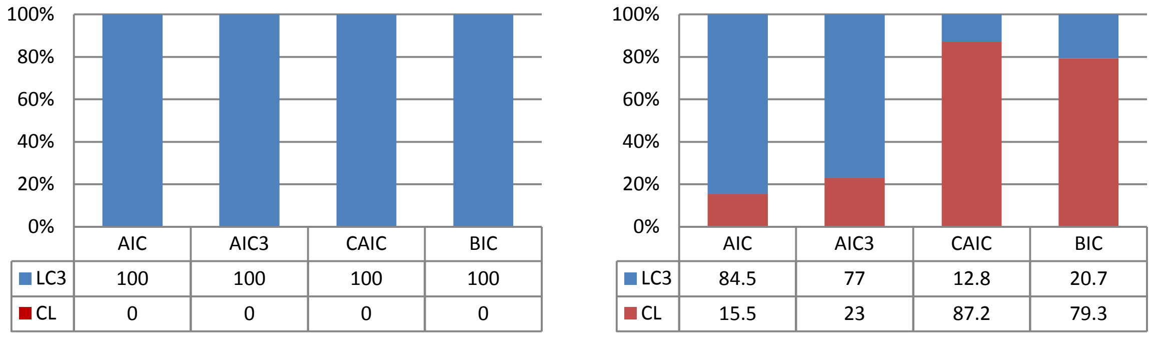


Fig. 8a. Lower variance scenario

| | AIC | AIC3 | CAIC | BIC |
|---|---|---|---|---|
| 🟦 LC3 | 100 | 100 | 100 | 100 |
| 🟥 CL | 0 | 0 | 0 | 0 |

*Note: N. of valid replications 787*

Fig. 8b. Higher variance scenario

| | AIC | AIC3 | CAIC | BIC |
|---|---|---|---|---|
| 🟦 LC3 | 84.5 | 77 | 12.8 | 20.7 |
| 🟥 CL | 15.5 | 23 | 87.2 | 79.3 |

*Note: N. of valid replications 430*

The relative performance of the MSC dramatically changes in the next exercises when the comparisons are between the LC-2 and LC-3 classes: Figures 9-10 report these comparisons. When the DGP is LC-2 (Figure 9), either in the lower and higher variance scenario, the AIC and AIC3 almost always select the wrong model, while BIC and especially CAIC perform very well.

**Fig. 9. DGP Latent 2 Classes: Latent 2 Classes vs Latent 3 Classes**

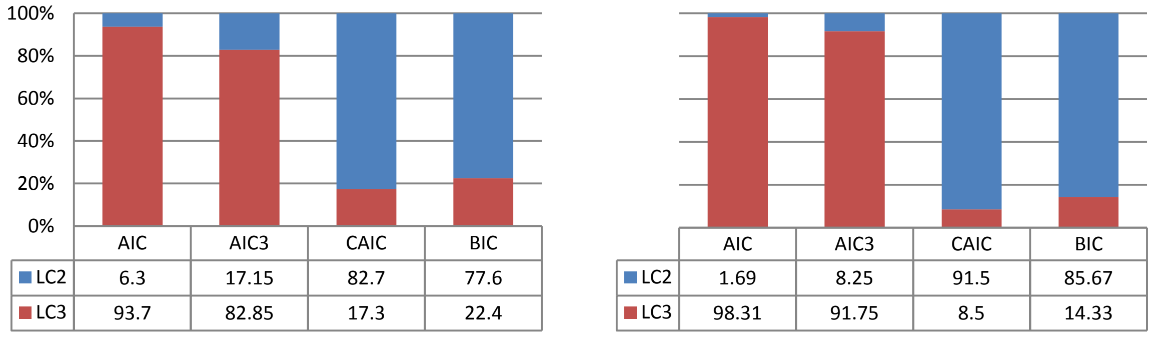


Fig. 9a. Lower variance scenario

| | AIC | AIC3 | CAIC | BIC |
|---|---|---|---|---|
| 🟦 LC2 | 6.3 | 17.15 | 82.7 | 77.6 |
| 🟥 LC3 | 93.7 | 82.85 | 17.3 | 22.4 |

*Note: N. of valid replications 746*

Fig. 9b. Higher variance scenario

| | AIC | AIC3 | CAIC | BIC |
|---|---|---|---|---|
| 🟦 LC2 | 1.69 | 8.25 | 91.5 | 85.67 |
| 🟥 LC3 | 98.31 | 91.75 | 8.5 | 14.33 |

*Note: N. of valid replications 824*

When the DGP is LC-3 (Figure 10), all criteria seem to work nicely in the lower variance setting, while in the higher variance setting the BIC and the CAIC underestimate the number of segments, respectively, in the 13% and 20% of cases. The tendency by AIC and AIC3 to overfit, and by CAIC and BIC to underfit the number of segments (especially when the choice data are more "noisy") confirms previous results by Andrews and Currim (2003) and Brochado and Martins (2006), as discussed in Section 2.

**Fig. 10. DGP Latent 3 Classes: Latent 3 Classes vs Latent 2 Classes**

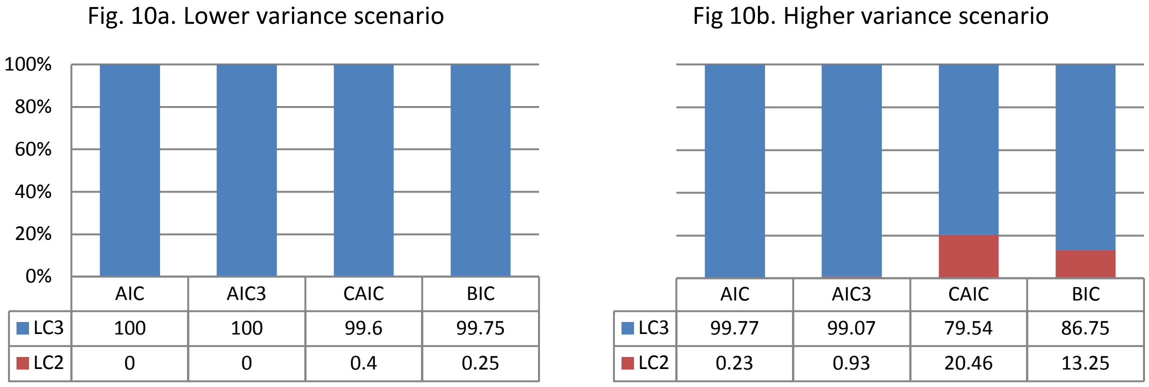


Fig. 10a. Lower variance scenario

| | AIC | AIC3 | CAIC | BIC |
|---|---|---|---|---|
| 🟦 LC3 | 100 | 100 | 99.6 | 99.75 |
| 🟥 LC2 | 0 | 0 | 0.4 | 0.25 |

*Note: N. of valid replications 787*

Fig 10b. Higher variance scenario

| | AIC | AIC3 | CAIC | BIC |
|---|---|---|---|---|
| 🟦 LC3 | 99.77 | 99.07 | 79.54 | 86.75 |
| 🟥 LC2 | 0.23 | 0.93 | 20.46 | 13.25 |

*Note: N. of valid replications 430*

The comparisons examined so far involved pairs of models, non-nested and “boundary” nested, characterized by different number of parameters. We have seen that in case of discrepancy, AIC and AIC3 tend to select the richer model, while BIC and CAIC select the more parsimonious specification; in comparing non nested models, the BAS and Horowitz tests resemble the AIC and AIC3 patterns, but with worse performance; while the Vuong test is most often unable to discriminate between models, even though when it does it is practically sure that it gives the correct indication.

Now we will compare non nested models with equal number of parameters. In this case, there is no difference among the MSC, and also the BAS and Horowitz tests reduce to the same formula. In particular, we contrast different specifications of the RPL model (Figures 11-13) where the five no price attributes are treated as random distributions while the price parameter is always kept fixed. Three scenarios are considered for this comparison. First, we consider the DGP RPL with normally distributed (RPL_N) parameters and we contrast the correct specification with that of a RPL model with parameters specified as log-normally distributed. Then we generate data from a DGP RPL model with log-normally distributed (RPL_LN) parameters, estimating the correct model and a RPL normally distributed parameters. Finally, we set the DGP so that three parameters are normally distributed and two are log-normally distributed (DGP mixed), estimating the correct model (i.e. RPL Mixed) and two RPL models, with respectively 5 normal and 5 log-normal random parameters.

Figure 11 reports the DGP RPL normal with comparison between RPL normal and log-normal or mixed normal and log-normal. Figure 11a and 12a show that the MSC select the correctly specified models in most comparisons between the Normal and the Log-normal specification.

**Fig. 11. DGP RPL_N: DGP RPL_N vs DGP RPL_LN and RPL_M**

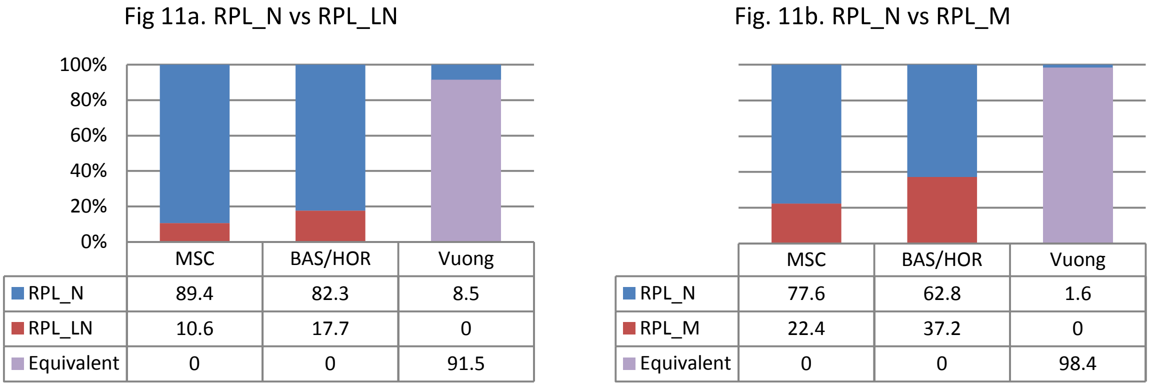


Fig 11a. RPL_N vs RPL_LN

| | MSC | BAS/HOR | Vuong |
|---|---|---|---|
| 🟦 RPL_N | 89.4 | 82.3 | 8.5 |
| 🟥 RPL_LN | 10.6 | 17.7 | 0 |
| 🟪 Equivalent | 0 | 0 | 91.5 |

*Note: N. of valid replications 955*

Fig. 11b. RPL_N vs RPL_M

| | MSC | BAS/HOR | Vuong |
|---|---|---|---|
| 🟦 RPL_N | 77.6 | 62.8 | 1.6 |
| 🟥 RPL_M | 22.4 | 37.2 | 0 |
| 🟪 Equivalent | 0 | 0 | 98.4 |

*Note: N. of valid replications 932*

When the comparison involves the Mixed specification model (Figure 11b and 12b) the performance of MSC declines although they are still preferable to the BAS/ Horowitz test. As regards the Vuong test, again in the majority of cases it cannot discriminate between models, even though, consistently with what previously noticed, it never selects the wrong model.

**Fig. 12. DGP RPL_LN: DGP RPL_LN vs DGP RPL_N and RPL_M**

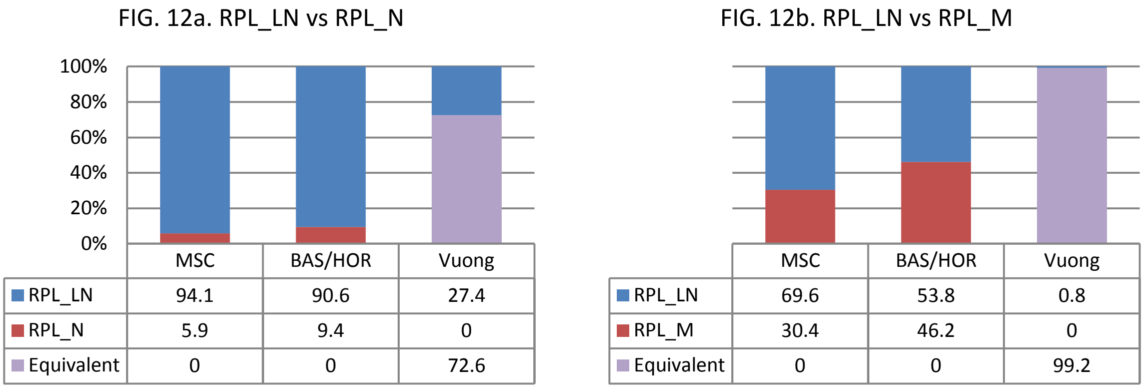


FIG. 12a. RPL_LN vs RPL_N

| | MSC | BAS/HOR | Vuong |
|---|---|---|---|
| 🟦 RPL_LN | 94.1 | 90.6 | 27.4 |
| 🟥 RPL_N | 5.9 | 9.4 | 0 |
| 🟪 Equivalent | 0 | 0 | 72.6 |

*Note: N. of valid replications 955*

FIG. 12b. RPL_LN vs RPL_M

| | MSC | BAS/HOR | Vuong |
|---|---|---|---|
| 🟦 RPL_LN | 69.6 | 53.8 | 0.8 |
| 🟥 RPL_M | 30.4 | 46.2 | 0 |
| 🟪 Equivalent | 0 | 0 | 99.2 |

*Note: N. of valid replications 932*

In Figure 13 we examine the case of DGP RPL mixed. The results echo previous MSC and tests performance: MSC select the correct specification more often than the BAS/HOR tests, and the Vuong test is generally unable to discriminate between models.

These results seem to suggest that the standard practice of choosing the distributional specification on the basis of improvements in the log-likelihood values is well supported.

**Fig. 13. DGP RPL_M: DGP RPL_LN vs DGP RPL_N and RPL_M**

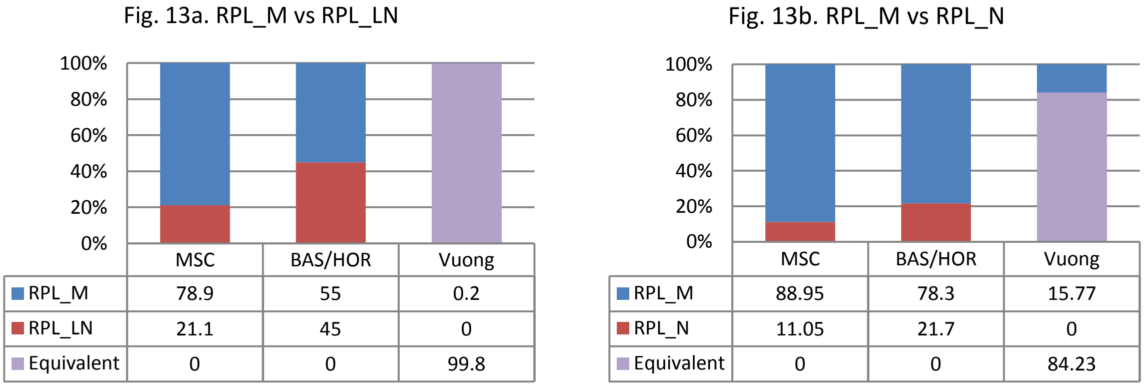


Fig. 13a. RPL_M vs RPL_LN

| | MSC | BAS/HOR | Vuong |
|---|---|---|---|
| 🟦 RPL_M | 78.9 | 55 | 0.2 |
| 🟥 RPL_LN | 21.1 | 45 | 0 |
| 🟪 Equivalent | 0 | 0 | 99.8 |

*Note: N. of valid replications 507*

Fig. 13b. RPL_M vs RPL_N

| | MSC | BAS/HOR | Vuong |
|---|---|---|---|
| 🟦 RPL_M | 88.95 | 78.3 | 15.77 |
| 🟥 RPL_N | 11.05 | 21.7 | 0 |
| 🟪 Equivalent | 0 | 0 | 84.23 |

*Note: N. of valid replications 507*

Finally, we report results of the analysis of the LR test performance in discriminating between nested models: the RPL and CL (see Table 2). It can be observed that when the DGP is CL the percentage of rejection of the null (CL) is smaller than the size of the test. When the DGP is RPL, the LR test leads to selection of the correctly specified model over the CL almost in every comparison when the DGP is lower variance; however the power of the test decreases sensibly in the experiment with higher variance.

**Table 4. Log-likelihood ratio tests CL versus RPL, % RPL is preferred**

| DGP | Level of significance | Low variance | High variance |
|---|---|---|---|
| **RPL**[a] | α=10% | 100 | 87.66 |
| | α=5% | 99.89 | 78.47 |
| | α=1% | 99.47 | 59.3 |
| **CL**[b] | α=10% | 3.37 | 1.93 |
| | α=5% | 0.76 | 0.74 |
| | α=1% | 0.15 | 0 |

[a] *RPL DGP: lower variance: 954 valid replications; higher variance: 892 valid replications.*

[b]

## 5. Empirical study

In this section we report an illustrative environmental economics CE study, where the MSC and tests previously analyzed can be implemented. The survey is detailed in Strazzera et al. (2012) and refers to a study on public acceptance of wind energy projects. Respondents, interviewed face-to-face by trained enumerators, valued six pairs of wind projects characterized by different plant locations, public benefits, ownership of the plant and private benefits in the form of energy bill savings. The designed attributes are reported in Table B in Appendix.

The aim of the study was to inform decision-making on the factors that may promote or hinder public acceptance of green energy projects, with a special attention on taking into account individuals' preference heterogeneity. Therefore different model specifications were estimated: the

CL model, the RPL with different distributional specifications for the random parameters, and several latent class models (LC with 2, 3, 4 classes, with and without covariates in the class probability model). The empirical data supports a LC model with four covariates in the segment membership probability model, which was selected through LR and Wald tests, and up to four segments. The LC models estimated on the empirical data are richer than the simple models used in our simulations, but what is most relevant for extrapolation of the Monte Carlo analysis results is that the statistical structure of the models is the same. Table 5 reports results for the best specifications of each model.

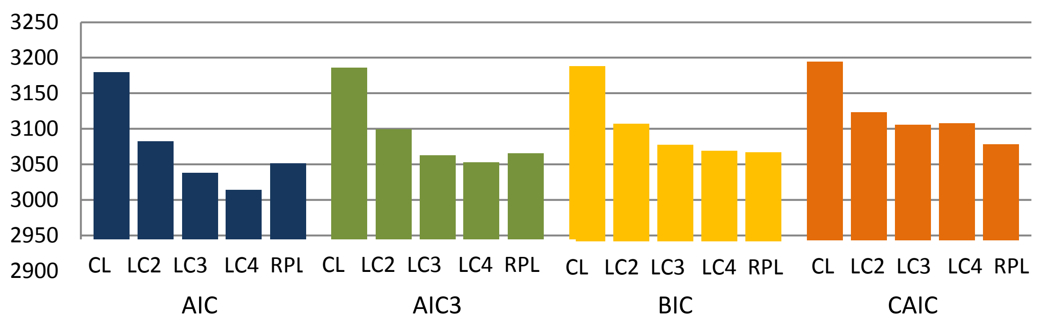
**Table 5. Conditional Logit, Random Parameters (Normal) Logit and Latent class models**

| Variable | **CL** | **RPL** | | **LC 4** | | | |
|---|---|---|---|---|---|---|---|
| | Coeffs<br>(Std.Err) | Coeffs<br>(Std.Err) | St.dev<br>(Std.Err) | Coeffs_Cl1<br>(Std.Err) | Coeffs_Cl2<br>(Std.Err) | Coeffs_Cl3<br>(Std.Err) | Coeffs_Cl4<br>(Std.Err) |
| Beach SI | 0.340***<br>(0.027) | 0.523***<br>(0.052) | 0.523***<br>(0.074) | 0.965***<br>(0.219) | 2.337***<br>(0.821) | 0.365***<br>(0.130) | 0.065<br>(0.062) |
| Beach MC | 0.476***<br>(0.034) | 0.680***<br>(0.067) | 0.798***<br>(0.080) | 2.034***<br>(0.324) | 0.589*<br>(0.321) | 0.368***<br>(0.135) | 0.245***<br>(0.091) |
| Arch_site | 0.339***<br>(0.046) | 0.560***<br>(0.080) | 0.768***<br>(0.120) | 1.188***<br>(0.311) | -0.597<br>(0.591) | -0.104<br>(0.217) | 0.671***<br>(0.107) |
| Property | 0.090***<br>(0.030) | 0.173***<br>(0.051) | 0.342***<br>(0.105) | -0.0638<br>(0.174) | 0.521*<br>(0.279) | -0.222<br>(0.152) | 0.300***<br>(0.079) |
| Services | 0.162***<br>(0.028) | 0.204***<br>(0.043) | 0.365***<br>(0.080) | 0.122<br>(0.123) | 0.254<br>(0.175) | 0.832***<br>(0.203) | -0.017<br>(0.069) |
| Bill saving | 1.677***<br>(0.145) | 2.276***<br>(0.216) | | 0.473<br>(0.896) | 3.297**<br>(1.504) | 5.554***<br>(1.209) | 1.120***<br>(0.301) |
| | | | | *Class Probability Model* | | | |
| Constant | | | | -0.764<br>(0.842) | -2.321**<br>(1.077) | -1.008<br>(1.137) | 0 |
| ID_SI Beach | | | | -0.073<br>(0.225) | 0.921***<br>(0.303) | 0.574*<br>(0.304) | 0 |
| ID_MC Beach | | | | 0.558**<br>(0.227) | -0.283<br>(0.277) | -0.980***<br>(0.328) | 0 |
| Consumerists | | | | 0.051<br>(0.214) | 0.357<br>(0.267) | 1.189***<br>(0.351) | 0 |
| Local Devoted | | | | -0.381*<br>(0.217) | -0.338<br>(0.270) | -0.831***<br>(0.295) | 0 |
| Average class probability | | | | 0.272 | 0.156 | 0.142 | 0.43 |
| Log likelihood | -1583.85 | -1514.75 | | -1435.28 | | | |
| LL(0) | -1796.63 | -1796.63 | | -1796.63 | | | |
| Pseudo R2 | 0.118 | 0.156 | | 0.201 | | | |
| Sample size | 432 | | | | | | |
| Observations | 2592 | | | | | | |

*** 1% significance; **5% significance; *10% significance

The baseline model is the CL, reported in the second column of the table: the model fit seems satisfactory, since all parameters are significant at 1% level, and the Chi-square test (not reported) indicates that the full model is better than the null. Now, we want to see whether the assumption of homogenous preferences of the CL model is tenable. Hence, we estimate a RPL model, and allow for heterogeneous preferences for all attributes, with the exclusion of the monetary attribute: the estimates of the utility coefficients and the standard deviation parameters are reported in columns 3 and 4 of the table. The examination of the significance of the standard deviation parameters (in this case all of them are significant at 1% level), suggests that there is indeed preference heterogeneity across individuals. In addition, we apply the LR test to check whether the RPL is significantly better than the CL model. The null hypothesis that the restricted specification (CL) is the true model is rejected at α=1% significance level, hence the RPL is preferred to the CL model. We may also want to compare the CL model versus the Latent Class models. In this case, we base our judgment on MSC. From the results of the Monte Carlo study, we learn that when the DGP is LC, either 2 or 3 classes, all penalization criteria work equally well, whereas when the DGP is CL, the CAIC and BIC are preferable. In Figure 14 the CL is compared with three LC models and the RPL model: the CL has the worst (highest) value for all penalization criteria.

Fig. 14. The performance of MSC in the empirical study



Clarified that the CL can be discarded, the researcher is left with the task of selecting the “best” specification among our competing models. Among the set of LC models, the LC-4 classes is selected by the MSC with AIC, AIC3 and BIC penalizations, whereas the CAIC points towards the LC-3 classes. In the comparison of LC models with the RPL model, we use the tests for non nested models in addition to the MSC for model selection; Table 6 summarizes main results.

**Table 6. RPL versus LC models, indications from MSC and Tests**

| Models | Model Selection Criteria | | | | BAS[a] | Horowitz[a] | Vuong[a] | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AIC | AIC3 | CAIC | BIC | | | AIC | AIC3 | CAIC | BIC |
| RPL | | | ● | ● | | | ● | ● | ● | ● |
| LC-4 | ● | ● | | | ● | ● | ● | ● | | |
| RPL | | | ● | ● | | | ● | ● | ● | ● |
| LC-3 | ● | ● | | | ● | ● | ● | ● | ● | ● |
| RPL | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| LC-2 | | | | | | | ● | ● | | ● |

*Note: a) is for α=5%*

Comparing with MSC the LC models with the RPL we find that AIC and AIC3 point toward the LC-4 classes whereas CAIC and BIC favor the RPL. The BAS and Horowitz test indicate that the LC-4 and LC-3 are significantly better than the RPL, while the latter is significantly better than the LC-2 classes. The Vuong test indicates that 1) the RPL is selected with the CAIC correction over the LC-2 classes, while the other penalizations do not discriminate; 2) RPL and LC-3 classes are valued equivalent by all penalizations; and finally, the RPL is chosen over the LC-4 classes by the CAIC penalization at α=1% significance level and by the BIC at α=5%.

Summing up, the LC-4 classes model is selected by the AIC and AIC3 MSC, and by the Horowitz and the Ben-Akiva and Swait tests; while the RPL is selected by the CAIC and BIC MSC, and by the Vuong-CAIC and BIC.

The results of the Monte Carlo study have shown that for competing models with different number of parameters, if the models are relatively close in terms of fit there will be discrepancy between the MSC: BIC and CAIC tend to choose the most parsimonious model, while AIC and AIC3 tend to choose the richer model. Although CAIC and BIC always gave correct indications in the lower variance scenarios, this was not true in some of the experiments with high variance disturbances. Since in empirical applications we do not know the true DGP, we still are left with uncertainty on what to choose. The indications by the BAS and the Horowitz tests seem not reliable, given their general bad performance in our Monte Carlo study; however we have seen that the Vuong test, although often inconclusive, virtually never selects a wrong model, and if it gives an indication, it can be quite safely trusted.  Hence, our model selection procedure indicates that the RPL model should be preferred to the CL and the LC specifications applied in this research.

## 6. Conclusions

The paper contributes to the discussion about the model selection criteria and tests for choice experiment data. A Monte Carlo analysis was designed to study the performance of model selection methods in comparing commonly used CE models: Conditional Logit, Random Parameters Logit and Latent Class. The attention is focused on some model selection criteria and tests that are based on the Kullback-Leibler Information Criterion. Some of these methods are popular in the environmental economics literature: the AIC, AIC3, CAIC and BIC MSC and the Ben-Akiva and Swait test; in addition, we include in the analysis also the Horowitz and the Vuong tests, which are rarely seen in choice experiment applications, but that could be useful in testing non nested models like Random Parameters Logit models vs Latent Class models.

Our results show that the MSC have different performances in different scenarios, but in general we could see that the AIC criterion, which is still the most common selection method employed in empirical applications, is outperformed by other criteria in almost all exercises. The BIC and especially CAIC outperform the other two criteria in all lower variance scenarios; however, in the case of DGP LC-3 classes with higher variance the opposite holds: the BIC and CAIC wrongly select the more parsimonious model (the bad performance is particularly serious when the LC-3 classes model is compared to the CL). It is possible that in this setting, the error term added to the three classes model makes the segmentation more "fuzzy", so that a simplified model (RPL or CL) is evaluated by the parsimonious criteria CAIC and BIC as fitting the data better than the correctly specified LC-3 classes model. This finding is noteworthy, since Keane and Wasi (2013, p.1043) argue that *"the number of segments identified by LC, and how sharply they differ, is a good preliminary diagnostic to use to determine the complexity of heterogeneity in a dataset"*. Depending on the criterion used, it may happen that the number of segments is heavily overestimated (by AIC and AIC3) or underestimated (by BIC and CAIC). Summing up, the results of our Monte Carlo study indicate that CAIC and BIC have similar properties, and the same holds for AIC3 and AIC; and that in general CAIC proved to be the most reliable criterion, and AIC the worst.

The performance of the tests in selecting between non nested models is quite disappointing. The BAS and the Horowitz tests are always outperformed by the MSC, and in general they make too often the wrong decision, especially in the exercises with RPL DGP; both tests seem quite unreliable, but the BAS test is even worse than the Horowitz. The performance of the Vuong test is also very unsatisfactory, since in most cases it cannot discriminate between models. On the other hand, and consistently with previous findings, it virtually never gives a wrong indication; hence, in spite of its very low power, it is worthwhile to apply this test to select between competing non nested models, since in the lucky event that it is able to discriminate between models we can be quite confident that it gives the correct indication.

The LR test works very well in selecting between the RPL and CL models in the lower variance setting, while in the high variance setting the power of the test declines, and the CL is mistakenly selected against the RPL in a significant number of comparisons (this situation resembles that of the CAIC and BIC criteria in selecting between LC-3 classes and CL models in the higher variance scenario, although it is certainly less serious); the empirical size of the test is always smaller than the nominal size. Since the t-statistic test on the standard deviation coefficients of the RPL have been shown by Mariel et al. (2013) to have higher empirical size than nominal, it would seem that a LR test can be used as a convenient complement to the check of standard deviation significance to inform a better model selection decision between these two models.

In conclusion, this study points out the importance of complementing the practitioner’s own judgment with the appropriate MSC or test to select the preferred specification. While we cannot suggest generalizing the results of our Monte Carlo analysis to models characterized by different statistical structures, we believe that our results shed some light on pros and cons of model selection procedures currently available for practitioners in the CE field.

In addition, our findings encourage researchers in designing further simulation studies for improving knowledge on the performance on model selection criteria and tests. To better understand the role of MSC and tests in selecting choice experiment models within the family of estimators considered in this paper, we suggest that further research be devoted to MSC and tests when the number of parameters, number of alternatives, number of respondents etc. vary systematically according to a specific study design. Finally, as more complex estimators are proposed and applied in the CE field, further studies should be designed to account for these different model structures.

## Notes

[1] Originally, the term Multinomial Logit was referred to models where the choice is conditional on socio-economic covariates and Conditional Logit to models conditioning on attribute covariates. In practice, the models are statistically equivalent, and the terms are used interchangeably.

[2] Papers selected among the top tier journals, from 2009 onwards, considering applied studies only.

[3] The Vuong test is characterized by good empirical size but low power (i.e. it rarely chooses the wrong model, but it often cannot discriminate between models): see Genius and Strazzera (2002) and Clark (2007); while the Clark test has been shown to have higher power, but at the price of selecting more often the wrong model (Clark 2007).

[4] elsa.berkeley.edu/Software/abstracts/train0296.html

[5] General reference on information measures can be found in Gourieroux and Monfort (1995).

[6] For example, Brochado and Martins (2006) and Mariel et al. (2013) use 500 replications.

## Appendix

**TABLE A**. CE studies quoted in Section 2

| Reference | Journal | Alternatives | Attributes | N. Levels (N. attributes) | N. choice sets | Sample size | Obs. | Models Estimated | Model Selection Methods |
|---|---|---|---|---|---|---|---|---|---|
| **Nested models** | | | | | | | | | |
| Amador et al. (2013) | EnE | 3 | 5 | 3(4) 2(1) | 9 | 376 | 3384 | CL/PML-EC | AIC/BIC/CAIC/$R^2$ |
| Czajkowski et al. (2013) | REnE | 3 | 3 | 3(2) 2(1) | 6 | 311 | 1371 | CL/G-MNL | AIC/LL/ $R^2$ |
| Lanz and Provins (2013) | ERE | 3 (1 status quo) | 4 | 3(4) 5(1) | 4 | 106 | 424 | CL/RPL | AIC/ $R^2$/SD |
| Nguyen et al. (2013) | EcE | 2 | 4 | 3(3) 2(1) | 6 | 1014 | 6084 | CL/RPL | LR/SD/AIC/BIC |
| Zander et al. (2013) | EcE | 3 (1 status quo) | 6 | 3(3) 2(2) 5(1) | 6 | 200 | 1194/1248 | RPL/S-MNL/G-MNL | LL/$R^2$/AIC |
| Duke et al. (2012) | EcE | 4 (1 status quo) | 5 | 2(3) 4(1) 5(1) | n.a. | 664 | 3280 | CL/RPL | LR |
| Gelo and Koch (2012) | EcE | 3 (1 status quo) | 4 | 2(2) 4(1) 5(1) | 4 | 600 | 2400 | CL/RPL | LR/SD |
| Kawata and Watanabe (2012) | EcE | 3 (1 opt out) | 3 | 3(2) 5(1) | 3 | 700 | 2100 | CL/RPL | LR |
| Achtnicht (2011) | EcE | 2 | 7 | 3(4) 4(1) 2(2) | 12 | 379 | 4548 | CL/RPL | LR/SD |
| Adamowicz et al. (2011) | JEEM | 4 or 5 | 3 or 5 | 4(2) or 4(4) 5(1) | 4 | 1219 | 4876 | CL/RPL | SD |
| Grisolia and Willis (2011) | AppEcon | 5 | 7 | 2(1) 3(1) 4(3) 5(2) | 12 | 332 | 3984 | CL/RPL | SD/ $R^2$ |
| Juutinen et al.(2011) | EcE | 3 | 5 | 3(4) 5(1) | 4 | 473 | 1892 | CL-RPL | SD |
| Onozaka and McFadden (2011) | AJAE | 3 (1 opt out) | 5 | 2(2) 3(3) | 8 | 629/554 | 5032 /4432 | CL/RPL | LR/SD |
| Susaeta et al. (2011) | EnE | 2 | 3 | 6(2) 4(1) | 6 | 182 | 1092 | CL-RPL | LR/SD |
| Ward et al. (2011) | EnE | 4 (1 opt out) | 6 | 4(4) 2(2) | 14 | 388 | 4732 | CL-RPL | LR |
| Van Loo et al. (2011) | FQP | 3 (1 opt out) | 2 | 3(1) 4(1) | 12 | 976 | 11712 | CL/RPL | LR |
| Abdullah and Mariel (2010) | EP | 3 (1 opt out) | 4 | 5(1) 2(1) 4(2) | 4 | 202 | 808 | CL/RPL | LR |
| Borg and Scarpa (2010) | ECE | 3 (1 opt out) | 9 or 6 | 3 | 9 or 8 | 86/ 198 | 2193/5346 | CL/RPL | AIC/AIC3/BIC |

**TABLE A**. Continued

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Jensen et al. (2010) | EnE | 4 | 8 | 5(1) 4(2) 3(1) 2(4) | 14 | 914 | 12796 | CL-RPL | LR |
| Kosenius (2010) | EcE | 3 (1 opt out) | 5 | 7(1) 3(4) | 6 | 726 | 3946 | CL/RPL | LR |
| Westerberg et al. (2010) | EcE | 3 (1 status quo) | 6 | 6(1) 3(5) | 9 | 90 | 810 | CL/RPL | SD |
| Dimitropoulos et al. (2009) | EP | 3 (1 opt out) | 5 | 2(3) 4(2) | 8 | 212 | 1696 | CL/RPL | LR/SD |
| Gracia et al. (2009) | FQP | 3 (1 opt out) | 4 | 2 | 4 | 400 | 3200 | CL/RPL | LR |
| Kataria (2009) | EnE | 3 (1 opt out) | 5 | 2 (1) 3(3) 7 (1) | 4 | 568 | 2222 | CL/RPL | LR |
| Shen (2009) | AppEcon | 3 | 5 | 2 | 8 | 467/453 | 3736/3624 | CL/RPL | LL |
| Birol et al. (2006) | EcE | 3 (1 opt out) | 5 | 2(3) 4(2) | 8 | 407 | 3256 | CL/RPL | LR |
| Greene and Hensher (2003) | TR B | 4(1 status quo) | 6 | 4 | 16 | 274 | 4384 | CL/RPL | LR/LL |
| | | | | **Boundary nested models** | | | | | |
| Czajkowski et al. (2013) | REnE | 3 | 3 | 3(2) 2(1) | 6 | 311 | 1371 | CL-LC | AIC/LL/$R^2$ |
| Kanchanaroek et al. (2013) | EcE | 3 (1 opt out) | 6 | 2(4) 4(2) | 10 | 272 | 2720 | CL/LC | AIC/BIC/ $R^2$ |
| Nguyen et al. (2013) | EcE | 2 | 4 | 3(3) 2(1) | 6 | 1014 | 6084 | CL/LC | AIC/BIC/ $R^2$ |
| Garrod et al. (2012) | EcE | 2 | 5 | 2 | 4 | 1397 | 4720 | CL/LC | AIC |
| Gelo and Koch (2012) | EcE | 3 (1 status quo) | 4 | 2(2) 4(1) 5(1) | 4 | 600 | 2400 | CL/LC | LR |
| Hensher et al. (2012)[a] | TR | 3 | / | / | 16 | / | / | CL/LC | BIC |
| Strazzera et al. (2012) | EP | 2 | 6 | 2(1) 2(4) 5(1) | 6 | 432 | 2592 | CL /LC | AIC/BIC/ $R^2$ |
| Thiene et al. (2012) | JFE | 3 (1 opt out) | 5 | 7(1) 2(1) 3(3) | 6 | 306 | 1836 | CL/LC | AIC/AIC3/BIC/CAIC |
| Hidrue et al. (2011) | REnE | 4 (1 status quo & 1 opt out) | 6 | 8(1) 4(5) | 2 | 3029 | 6058 | CL/LC | AIC/BIC |
| Borg and Scarpa (2010) | EcE | 3 (1 opt out) | 9 or 6 | 3 | 9 or 8 | 86 /198 | 2193/5346 | CL/LC | AIC/AIC3/BIC |
| Brouwer et al. (2010) | LE | 3(1 status quo) | 4 | 2(1) 3(1) 4(1) 6(1) | 4 | 619 | 2476 | LC | AIC/BIC |
| Dietz and Atkinson (2010) | LE | 4(1 status quo) | 4 | 3(2) 2(1) 9(1) or (2(2) 3(1) 9(1) | 10 | 231 or 237 | 2310 or 2370 | LC | AIC/BIC/ $R^2$ |
| Meyerhoff et al. (2010) | EP | 3 | 5 | 5(1) 3(4) | 5 | 708 | 3540 | CL/LC | AIC/BIC/CAIC |

**TABLE A**. Continued

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Kosenius (2010) | EcE | 3 (1 opt out) | 5 | 7(1) 3(4) | 6 | 726 | 3946 | CL/LC | AIC/BIC |
| Burton and Rigby (2009) | ERE | 3 | 4 | 7(1) 3(1) 4(2) | 9 | 608 | 5472 | CL/LC | AIC/AIC3/BIC<br>CAIC |
| Colombo et al. (2009) | AgE | 3 (1 status quo) | 6 | 3(5) 6(1) | 6 | 300 | 1187 | LC | AIC/CAIC |
| Shen (2009) | AppEcon | 3 | 5 | 2 | 8 | 467/453 | 3736/3624 | CL/LC | AIC/CAIC |
| Shen and Saijo (2009) | JEnvM | 4 | 6 | n.a. | 6 | 600 | 3600 | CL/LC | AIC/AIC3/BIC |
| Hynes et al. (2008) | AJAE | / | 6 | / | / | 279 | / | CL/LC | AIC/AIC3/BIC<br>crAIC |
| Birol et al. (2006) | EcE | 3 (1 opt out) | 5 | 2(3) 4(2) | 8 | 407 | 3256 | CL/LC | AIC/BIC |
| Greene and Hensher (2003) | TR B | 4 (1 status quo) | 6 | 4 | 16 | 274 | 4384 | CL/LC | LL |
| Boxall and Adamowicz (2002) | ERE | 6 (1 status quo) | 5 | 4 | 8 | 620 | 4892 | LC | AIC/BIC |
| | | | | **Non nested models** | | | | | |
| Broch and Vedel (2013) | ERE | 3 (1 status quo) | 4 | 3(4) | 6 | 853 | 5053 | RPL/LC | AIC/BIC/ R$^2$ |
| Czajkowski et al. (2013) | REnE | 3 | 3 | 3(2) 2(1) | 6 | 311 | 1371 | G-MNL/LC | AIC/LL/ R$^2$ |
| Gelo and Koch (2012) | EcE | 3 (1 status quo) | 4 | 2(2) 4(1) 5(1) | 4 | 600 | 2400 | RPL/LC | LL |
| Hensher et al. (2012)[a] | TR | 3 | / | / | 16 | / | / | LC/RPL/G-MNL | BIC |
| Keane and Wasi (2012)[b] | JAE | / | / | / | / | / | / | LC/RPL/G-MNL/MM-MNL | BIC |
| Tesfaye and Brouwer (2012) | EcE | 3 (1 opt out) | 6 | 2 (1) 3(1) 4(1) 5(1) 6(1) | 9 | 750 | 6750 | RPL/LC | AIC/BIC/ R$^2$ |
| Borg and Scarpa (2010) | EcE | 3 (1 opt out) | 9 6 | 3 | 9 | 86 198 | 2193/5346 | RPL/LC | AIC/AIC3/BIC |
| Brouwer et al. (2010) | LE | 3(1 status quo) | 4 | 2(1) 3(1) 4(1) 6(1) | 4 | 619 | 2476 | LC/RPL | BAS |
| Kosenius (2010) | EcE | 3 (1 opt out) | 5 | 7(1) 3(4) | 6 | 726 | 3946 | RPL/LC | BAS |
| Burton and Rigby (2009) | ERE | 3 | 4 | 7(1) 3(1) 4(2) | 9 | 608 | 5472 | RPL/DHCL/DHRPL/LC | Clark[c] |

**TABLE A**. Continued

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Colombo et al. (2009) | AgE | 3 (1 status quo) | 6 | 3(5) 6(1) | 6 | 300 | 1187 | LC/RPL/CH | BAS |
| Shen (2009) | AppEcon | 3 | 5 | 2 | 8 | 467/453 | 3736/3624 | RPL/LC | BAS |
| Birol et al. (2006) | EcE | 3 (1 opt out) | 5 | 2(3) 4(2) | 8 | 407 | 3256 | RPL/LC | BAS |

[a]7 data sets used. [b]10 data sets used. [c]Clark (2003). **SD**: significance of standard deviations. **CH**: Covariance Heterogeneity model . **EC**: Error Components model. **PML-EC**: Panel Mixed Logit with error component. **MNP**: Multinomial Probit. **DHCL**: Double Hurdle conditional Logit. **DHRPL**: Double Hurdle RPL**. S-MNL**: Scale-Multinomial Logit. **G-MNL**: Generalized Multinomial Logit. **MM-MNL**: Mixed-mixed MNL.

**Journal abbreviations= AgE**: Agricultural Economics. **AJAE**: American Journal of Agricultural Economics. **AppEcon**: Applied Economics. **EcE:** Ecological Economics. **EnE**: Energy Economics. **EP**: Energy Policy. **ERE**: Environmental and Resource Economics. **FQP**: Food Quality and Preference. **JAE**: Journal of Applied Econometrics. **JEM**: Journal of Environmental Management. **JEEM**: Journal of Environmental Economics and Management. **JFE**: Journal of Forest Economics. **LE**: Land Economics. **REnE**: Resource and Energy Economics. **TR B**: Transportation Research B. **TR**: Transportation.

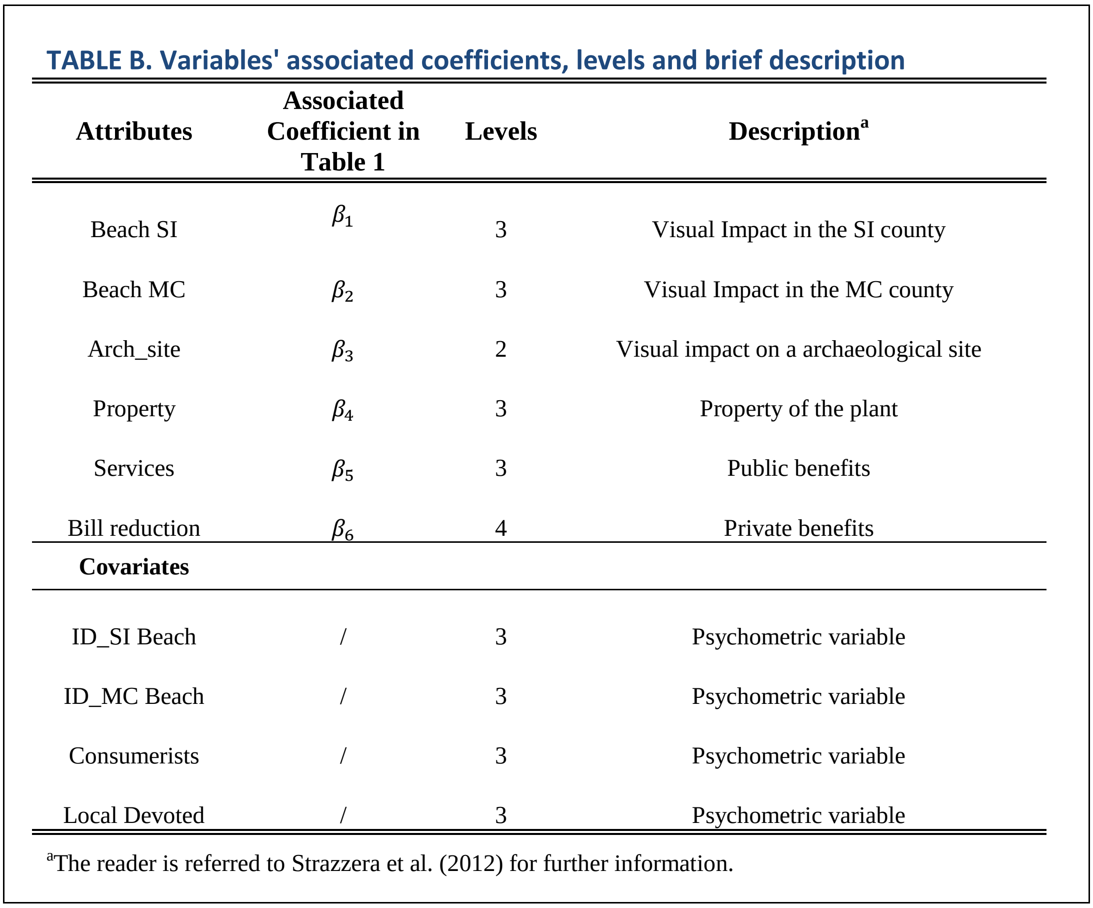


TABLE B. Variables' associated coefficients, levels and brief description

| Attributes | Associated Coefficient in Table 1 | Levels | Description[a] |
|---|---|---|---|
| Beach SI | $\beta_1$ | 3 | Visual Impact in the SI county |
| Beach MC | $\beta_2$ | 3 | Visual Impact in the MC county |
| Arch_site | $\beta_3$ | 2 | Visual impact on a archaeological site |
| Property | $\beta_4$ | 3 | Property of the plant |
| Services | $\beta_5$ | 3 | Public benefits |
| Bill reduction | $\beta_6$ | 4 | Private benefits |
| **Covariates** | | | |
| ID_SI Beach | / | 3 | Psychometric variable |
| ID_MC Beach | / | 3 | Psychometric variable |
| Consumerists | / | 3 | Psychometric variable |
| Local Devoted | / | 3 | Psychometric variable |

[a]The reader is referred to Strazzera et al. (2012) for further information.

## References

Abdullah S, Mariel P (2010) Choice experiment study on the willingness to pay to improve electricity services. Energy Policy 38:4570-4581

Achtnicht M (2011) Do environmental benefits matter? Evidence from a choice experiment among house owners in Germany. Ecological Economics 70:2191-2200

Adamowicz W, Dupont D, Krupnik A (2011) Valuation of cancer and microbial disease risk reductions in municipal drinking water: An analysis of risk context using multiple valuation methods. Journal of Environmental Economics and Management 61:213-226.

Amador F J, González R M, Ramos-Real, F J (2013) Supplier choice and WTP for electricity attributes in an emerging market: The role of perceived past experience, environmental concern and energy saving behavior. Energy Economics (in press)

Andrews R L, Currim, I S (2003) A comparison of segment retention criteria for finite mixture Logit models. Journal of Marketing Research 40(2): 235-243

Akaike H (1987) Factor analysis and AIC. Psychometrika 52:317-332

Ben-Akiva M, Swait J (1986) The Akaike likelihood index. Transportation Science 20:133-136.

Bhat C R (1997) An endogenous segmentation mode choice model with an application to intercity travel. Transportation Science 31:34-48

Birol E, Karousakis K, Koundouri P (2006) Using a choice experiment to account for preference heterogeneity in wetland attributes: The case of Cheimaditida wetland in Greece. Ecological Economics 60:145-156

Borg N B, Scarpa R (2010) Valuing quality changes in Caribbean coastal waters for heterogeneous beach visitors. Ecological Economics 69:1124-1139

Boxall P C, Adamowicz W L (2002) Understanding heterogeneous preferences in random utility models: A latent class approach. Environmental and Resource Economics 23:421-446

Bozdogan H (1987) Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika 52:345-370

Bozdogan H (1994) Proceedings of the first US/Japan conference on the frontiers of Statistical modeling: An informational approach. Kluwer Academic Publishers Dordrecht

Broch S W, Vedel S E (2012) Using choice experiments to investigate the policy relevance of heterogeneity in farmer Agri-Environmental contract preferences. Environmental and Resource Economics 51:561-581

Brochado A O, Martins F V (2006) Examining the segment retention problem for the "Group Satellite". FEP Working Paper N. 220, Universidade do Porto

Brouwer R, Martin-Ortega J, Berbel J (2010) Spatial preference heterogeneity: A choice experiment. Land Economics 86(3):552-568

Burton M, Rigby D (2009) Hurdle and Latent Class approaches to serial non-participation in choice models. Environmental and Resource Economics 42:211-226

Carlsson F, Martinsson P (2008) Does it matter when a power outage occurs? A choice experiment study on the willingness to pay to avoid power outages. Energy Economics 30:1232-1245

Clark K A (2003) Nonparametric model discrimination in international relations. Journal of Conflict Resolution 47(1):72-93

Clark K A (2007) A simple distribution free test for non-nested hypothesis. Political Analysis 15:347-363

Colombo S, Hanley N, Louviere J J (2009) Modelling preference heterogeneity in stated choice data: an analysis for public goods generated by agriculture. Agricultural Economics 40:307-322

Czajkowski M, Kądziela T, Hanley N (2013) We want to sort! Assessing households' preferences for sorting waste. Resource and Energy Economics (in press)

Czajkowski M, Buszko-Briggs M, Hanley N (2009) Valuing changes in forest biodiversity. Ecological Economics 68:2910-2917

Czajkowski M, Scasny M (2010) Study on benefit transfer in an international setting. How to improve welfare estimates in the case of the countries' income heterogeneity? Ecological Economics 69:2409-2416

Diez S, Atkinson G (2010) The equity-efficiency trade off in environmental policy: Evidence from stated preferences. Land Economics 86(3):423-443

Dimitropoulos A, Kontoleon A (2009) Assessing the determinants of local acceptability of wind farm investment: a choice experiment in the Greek Aegean Islands. Energy Policy 37:1842-1854

Duke J M, Borchers A M, Johnston R J, Absetz S (2012) Sustainable agricultural management contracts: Using choice experiments to estimate the benefits of land preservation and conservation practies. Ecological Economics 74:95-103

Fosgerau M, Bierlaire M (2007) A practical test for the choice of mixing distribution in discrete choice models. Transportation Research Part B: Methodological 41(7):784-794

Garrod G, Ruto E, Willis K, Powe N (2012) Heterogeneity of preferences for the benefits of environmental stewardship: A latent class approach. Ecological Economics 76:104-111

Gelo D, Koch S (2012) Does one size fit all? Heterogeneity in the valuation of community forestry programs. Ecological Economics 74:85-94

Genius M, Strazzera E (2011) Can unbiased be tighter? Assessment of methods to reduce the bias-variance trade-off in WTP estimation. Resource and Energy Economics 33(1):293-314

Genius M, Strazzera E (2002) A note on model selection and tests for non-nested contingent valuation models. Economics Letters 74:363-370

Genius M, Strazzera E (2001) Model Selection and Tests for Non Nested Contingent Valuation Models: An Assessment of Methods. FEEM Working Paper No. 34.2001

Glenk K, Fischer A (2010) Insurance, prevention or just wait and see? Public preferences for water management strategies in the context of climate change. Ecological Economics 69:2279-2291

Gourieroux C, Monfort A (1995) Statistics and Econometric Models. Cambridge University Press, Cambridge

Gracia A, Loureiro M L, Nayga Jr R M (2009) Consumers’ valuation of nutritional information: A choice experiment study. Food Quality and Preference 20:463-471

Greene W H, Hensher D A (2003) A latent class model for discrete choice analysis: contrasts with mixed Logit. Transportation Research B: Methodological 37:681-698

Grisolía J M, Willis K G (2011) An evening at the theatre: using choice experiments to model preferences for theatres and theatrical productions. Applied Economics 43:3987-3998

Halkos G E, Jones N (2012) Modeling the effect of social factors on improving biodiversity protection. Ecological Economics 78:90-99

Hensher D A, Greene W (2003) The mixed Logit model: the state of practice. Transportation 30:130-176

Hensher D A, Rose J, Li Z (2012) Does the choice model method and/or the data matter? Transportation 39:351-385

Hess S (2010) Conditional parameter estimates from Mixed Logit models: distributional assumptions and a free software tool. Journal of Choice Modelling 3:134-152

Hidrue M K, Parsons G R, Kempton W, Gardner M P (2011) Willingness to pay for electric vehicles and their attributes. Resource and Energy Economics 33:686-705

Horowitz J (1983) Statistical comparison of non nested probabilistic choice models. Transportation Science 17(3):319-350

Hynes S, Hanley N, Scarpa R (2008) Effects on welfare measures of alternative means of accounting for preference heterogeneity in recreational demand models. American Journal of Agricultural Economics 90:1011-1027

Jeffries N O (2003) A note on “Testing the number of components in a normal mixture”. Biometrika 90:991-994

Jensen K L, Clark C D, English B C, Menard R J, Skahan D K, Marra A C (2010) Willingness to pay for E85 from corn, switchgrass, and wood residues. Energy Economics 32:1253-1262

Juutinen A, Mitani Y, Mäntymaa E, Shoji Y, Siikamäki P, Svento R (2011) Combining ecological and recreational aspects in national park management: A choice experiment application. Ecological Economics 70:1231-1239

Kanchanaroek Y, Termansen M, Quinn C (2013) Property rights regimes in complex fishery management systems: A choice experiment application. Ecological Economics 93:363-373

Kataria M (2009) Willingness to pay for environmental improvements in hydropower regulated rivers. Energy Economics 31:69-76

Kawata Y, Watanabe M (2012) Valuing the mortality risk of wildlife reintroduction: Heterogeneous risk preferences. Ecological Economics 76:79-86

Keane M, Wasi N (2013) Comparing alternative models of heterogeneity in consumer choice behavior. Journal of Applied Econometrics 28:1018-1045

Kosenius A (2010) Heterogeneous preferences for water quality attributes: The case of eutrophication in the Gulf of Finland, the Baltic Sea. Ecological Economics 69:528-538

Kullback S, Leibler R A (1951) On Information and Sufficiency. Annals of Mathematical Statistics 22:79-86

Lanz B, Provins A (2013) Valuing local environmental amenity with discrete choice experiments: Spatial scope sensitivity and heterogeneous marginal utility of income. Environmental and Resource Economics 56:105-130.

Layton D F (1996) Rank-ordered, random coefficients multinomial probit models for stated preferences surveys. Paper presented at the 1996 Association of Environmental and Resource Economists Workshop, Talhoe City, California, 2-4 June 1996

Lo Y, Mendell N R, Rubin D B (2001) Testing the number of components in a normal mixture. Biometrika 88: 767-778

Louviere J J, Hensher D A, Swait J D (2000) Stated choice methods-Analysis and application. Cambridge University Press

Mariel P, de Ayala A, Hoyos D, Abdullah S (2013) Selecting random Parameters in discrete choice experiment for environmental valuation: A simulation experiment. Journal of Choice Modelling 7:44-57

Martinez-Cruz A L (2013) Comparing the performance of conditional, mixed, and latent class Logit models. EAERE 20th Annual Conference, Toulouse, 26-29 June 2013

McFadden D (1974) Conditional Logit analysis of qualitative choice behavior. In Frontiers in Econometrics. Ed. P. Zarembka, New York: Academic Press, 105-142

McFadden D, Train K (2000) Mixed MNL models for discrete responses. Journal of Applied Econometrics 15:447-470

Meijer E, Rouwendal J (2006) Measuring welfare effects in models with random coefficients. Journal of Applied Econometrics 21:227-244

Meyerhoff J, Ohl C, Hartje V (2010) Landscape externalities from onshore wind power. Energy Policy 38:82-92

Morey E R, Rowe R D, Watson M (1993) A Repeated Nested-Logit Model of Atlantic Salmon Fishing. American Journal of Agricultural Economics 75:578–592

Nguyen T C, Robinson J, Kaneko S, Komatsu S (2013) Estimating the value of economic benefits associated with adaptation to climate change in a developing country: A case study of improvements in tropical cyclone warning services. Ecological Economics 86:117-128

Onozaka Y, McFadden D T (2011) Does local labelling complement or complete with other sustainable labels? A conjoint analysis of direct and joint values for fresh produce claims. American Journal of Agricultural Economics 93(3):693-706

Ortùzar J de D, Eluru N, Srinivasan K K (2012) Methodological developments in activity-travel behavior. In Pendyala R M, Bhat C R (ed) Travel Behaviour Research in an Evolving World, p.357-366

Provencher B, Bishop R C (2004) Does accounting for preference heterogeneity improve the forecasting of a random utility model? A case study. Journal of Environmental Economics and Management 48:793-810

Revelt D, Train K (1998) Mixed Logit with repeated choices: households' choices of appliance efficiency level. Review of Economics and Statistics 80:647-657

Schwartz G (1978) Estimating the dimension of a model. The Annals of Statistics 6:461-464

Shen J (2009) Latent class model or mixed Logit model? A comparison by transport mode choice data. Applied Economics 41:2915-2924

Shen J, Saijo T (2009) Does an energy efficiency label alter consumers' purchasing decisions? A latent class approach based on a stated choice experiment in Shanghai. Journal of Environmental Management 90:3561-3573

Sillano M, Ortùzar J D (2005) Willingness to pay estimation with mixed Logit models: some new evidence. Environmental and Planning A 37:525-550

Spanos A (2010) Akaike-type criteria and the reliability of inference: Model selection versus statistical model specification. Journal of Econometrics 158:204-220

Strazzera E, Mura M, Contu D (2012) Combining choice experiments with psychometric scale to assess the social acceptability of wind energy projects: A latent class approach. Energy Policy 48:334-347

Strazzera E, Contu D, Ferrini S (2013) Check it out! A Monte Carlo analysis of the performance of selection criteria and tests for choice experiments models. International Choice Modeling Conference, Sydney 2013

Susaeta A, Lal P, Alavalapati J, Mercer E (2011) Random preferences towards bioenergy environmental externalities: A case study of woody biomass based electricity in the Southern United States. Energy Economics 33:1111-1118

Swait J (1994) A structural equation model of latent segmentation and product choice for cross-sectional revealed preference choice data. Journal of Retail and Consumer Services 1:77-89

Tesfaye A, Brouwer R (2012) Testing participation constraints in contract design for sustainable soil conservation in Ethiopia. Ecological Economics 73:168-178

Thiene M, Meyerhoff J, De Salvo M (2012) Scale and taste heterogeneity for forest biodiversity: Models of serial nonparticipation and their effects. Journal of Forest Economics 18:355-369

Torres C, Hanley N, Colombo S (2011) Incorrectly accounting for taste heterogeneity in choice experiments: Does it really matter for welfare measurement? Stirling Economics Discussion Papers 2011-02

Train K (2003) Discrete choice methods with simulations. Cambridge University Press

Train K (1998) Recreation demand models with taste variation over people. Land Economics 74:230-239

Tuma M N, Decker R (2013) Finite mixture models in market segmentation: A review and suggestions for best practies. Electronic Journal of Business Research Methods 11(1):2-15

Van Loo E J, Caputo V, Nayga Jr R M, Meullenet J F, Ricke S C (2011) Consumers' willingness to pay for organic chicken breast: Evidence from choice experiment. Food quality and Preference 22:603-613

Vuong Q H (1989) Likelihood ratio tests for model selection and non-nested hypothesis. Econometrica 57(2):307-333

Walker J L, Li J (2007) Latent lifestyle preferences and household location decisions. Journal of Geographical Systems 9:77-101

Ward D O, Clark C D, Jensen K L, Yen S T (2011) Consumer willingness to pay for appliances produced by Green Power Partners. Energy Economics 33:1095-1102

Westerberg V H, Lifran R, Olsen S B (2010) To restore or not? A valuation of social and ecological functions of the Marais des Baux wetland in Southern France. Ecological Economics 69:2383-2393

Wooldridge J M (2010) Econometric analysis of cross section and panel data. The MIT press, Cambridge

Zander K K, Signorello G, De Salvo M, Gandini G, Drucker A G (2013) Assessing the total economic value of threatened livestock breeds in Italy: Implications for conservation policy. Ecological Economics 93:219-229

# **Chapter 2:** Reliability and use in model estimation of follow up statements in choice modeling

# Reliability and use in model estimation of follow up statements in choice modeling[*]

## Abstract

We combine information on the stated most important attributes with Choice Experiment data, introducing a very simple strategy which leads to promising results. The information regarding the most important attribute is a subset of Stated Attribute Information, which consists in asking respondents to provide a full ranking of the attributes right after completing the choice experiment. The use of self-reported information has received criticisms in the choice modeling literature, an example being stated attribute attendance. Alternatively, complex econometric models have been proposed to infer attribute attendance, hence completely discarding follow up statements. However, these are not necessarily unreliable per se, as evidence has been repeatedly showing that what we treat as ignored may be actually only less important. In this paper, we show that another option can be considered, which proves to be simple yet effective: including in model estimation an alternative specific constant indicating which alternative contains the better level of the stated most important attribute.

**Keywords:** Stated Most Important Attribute·Stated Attribute Importance·Choice Experiments·Self-reported information

**Abbreviations**

[*] A preliminary version of this work has been submitted to the 5th World Congress of Environmental and Resource Economics (Contu and Strazzera 2014). I also acknowledge the work of Lidia Carboni, Roberta Mattana, Giulia Soru and Manuel Vinci with regard to data collection. Usual disclaimer applies.

| | |
|---|---|
| ANA | Attribute Stated Attendance |
| CE | Choice Experiment |
| CL | Conditional Logit |
| DTD | Door To Door |
| GA_NoTax | First component-PCA applied to Attributes’ ranking (Table 6) |
| LC | Latent Class |
| LI | Least Important (attribute) |
| PCA | Principal Component Analysis |
| PV | Photovoltaic |
| PV_NoIns | First component-PCA applied to Attributes’sub- ranking (Table 7) |
| ROL | Ranked Ordered Logit |
| RPL | Random Parameters Logit |
| SAI | Stated Attribute Importance |
| Sol_Str | Third component-PCA applied to Attributes’sub- ranking (Table 7) |
| ST | Solar Thermal |
| STR | Street Dumpsters |
| Tech_NoEm | Third component-PCA applied to Attributes’ ranking (Table 6) |
| THE | Thernal Insulation |
| TRA_NoEm | Second component-PCA applied to Attributes’ ranking (Table 6) |
| UD_NoDoor | Second component-PCA applied to Attributes’sub- ranking (Table 7) |
| UND | Underground Dumpsters |

## 1. Introduction

The choice modeling literature has acknowledged respondents might not adopt a fully compensatory behavior. Examples include lexicographic behavior (Sælensminde 2001), joint evaluation of common metric attributes (Hensher and Green 2010), attribute non-attendance (ANA) (Hensher et al. 2005). Results and policy implications may be distorted if the problem is not properly taken into account. As far as ANA is concerned, respondents might ignore one or more attributes when making the choice. To date, no consensus appears to be formed about the best strategy on 1) how to identify individuals who ignore some attribute and 2) how to model data when ANA has been detected. Regarding the former, two strategies have been used so far: a) asking respondents whether or not they ignored (stated or self-reported ANA) and b) inferring it by means of a given econometric framework (inferred ANA). Stated ANA was firstly proposed by Hensher et al. (2005), but it has been questioned on the grounds that the information obtained does not seem to be reliable (Campbell and Lorimer 2009; Carlsson et al. 2010; Hess and Hensher 2010; Hess 2012; Hess et al. 2013; Kragt 2013), an exception being Hole et al. (2013). Specifically, the common critique is that individuals may indicate as ignored what in reality has been only *less* important.

The importance of distinguishing between ignored and less important attributes has been emerging in the inferred ANA literature as well. For example, it has been suggested that attenuating the mean and the variance of the marginal utility represents a significant improvement over restricting the parameter to zero or treating as missing data those individuals more likely to have ignored some attributes (Carlsson et al. 2010; Balcombe et al. 2011; Cameron and DeShazo 2011; Kehlbacher et al. 2013). All in all, it seems unsatisfactory to simply divide respondents between attendants and non-attendants. In turn, as far as self-reported information is concerned, it may be inadequate to simply ask whether an attribute has been ignored or not. In this direction is the work of Colombo et al. (2013) and Scarpa et al. (2013), who asked respondents to state their frequency of attendance (e.g.: always, sometimes, never); while Alemu et al. (2013) ask to express the reasons why an attribute has been ignored.

In this work we follow a different strategy: we asked respondents to rank the attributes, from the most to the least important, right after the series of choice tasks. Additionally, an open ended question has been posed concerning the reason of the first and last place attached to the attributes. After the start of this project, we got aware of an unpublished work by Balcombe et al. (2012), where a ranking of the attributes was requested after the choice experiment and modeled in a mixed Logit framework following two strategies. Namely, using the ranking data as covariates and estimating a contraction factor for those individuals who ranked the attributes as the least important, finding significant improvements as opposed to the model without the ranking information.

We also acknowledge this follow-up question has been used in the past to assess internal consistency of choice experiment data (Azevedo et al. 2009). However, only with Balcombe et al. (2012) it was introduced the idea of using stated attribute importance (SAI) to complement choice experiment data in model estimation. In this paper we contribute in assessing SAI's reliability by means of internal consistency checks. An internal consistency check, based on the stated most important attributes information only, in turn gives birth to the best model according to our empirical application. Crucially, in the choice modelling literature there has been some interest towards Best-Worst scaling instead of asking the respondent to choose one alternative (Flynn 2010). In this paper, we are focusing on the use of a stated attribute ranking after the choice experiment, investigating its reliability and potential use in model estimation.

Summarizing, whilst assessing whether asking respondents to state attribute importance provides useful information and whether it might be preferable to stated ANA, we propose a simple yet effective strategy to include a subset of stated attribute attendance in model estimation. The outline of the paper is the following: in the next section we provide a brief overview of the internal consistency strategies employed; in section 3 the survey is introduced; section 4 describes the methodology; section 5 presents the SAI results; section 6 deals with the internal consistency checks and finally section 7 concludes.

## 2. Internal consistency checks

If we are able to effectively elicit respondents' preferences by means of multiple choice tasks, we would expect the implicit ranking derived from CE data to be consistent with the explicit ranking, obtained by directly asking individuals to state the importance of each of the attributes. As effectively put forward by Ryan and San Miguel (2000), '*if commodity A is preferred to B, then individuals should be willing to pay more for A than B'*. Hence, a simple test of internal consistency is to compute the willingness to pay (WTP) estimates and check whether those attributes that have been ranked as relatively more important are also associated with a greater willingness to pay. This is the strategy undertaken in Azevedo et al. (2009), who showed that mean willingness to pay proved to be consistent with the stated ranking.

Next, we focus on individual preferences, again comparing them to the rankings. First we estimate a Random Parameters Logit and check for the significant presence of heterogeneity in the data. Next, we compute the individual coefficients, conditioning previous RPL's results on observed choice. If there is consistency, we expect that the more (stated) important an attribute is, the greater the mean of its coefficient. In addition, we check whether high values of the ratio between individual standard deviations over individual coefficients' mean (Hess and Hensher 2010) are linked to less important attributes.

We further explore taste heterogeneity including variables derived from the ranking and including them into model estimation. We expect respondents who have ranked a given attribute as relatively more important to be associated with a greater magnitude for the correspondent coefficient. Conversely, we expect those who ranked a given attribute as relatively less important to be associated with a lower magnitude (or not significant one) for the correspondent coefficient. This strategy involves a two-stage procedure which is suggested only as internal reliability check.

We then consider segments of respondents and employ a common inferred-ANA approach, originally proposed by Scarpa et al. (2009): a confirmatory latent class model framework. Once computed the predicted probabilities of belonging to a given class for each respondent, we check the composition of the class in terms of stated attribute importance. For instance, if class $c$ is characterized by a parameter constrained to zero for attribute k, we then expect to find those who ranked attribute k as less important to be significantly associated to class *c*.

Finally, the stated most important attributes might also capture lexicographic behavior. Namely, the respondent may overly focus only on a given level of an attribute. If this is the case, the variable choice should be largely explained by a constant indicating which alternative contains that level; we call this constant $ASC_{Most}$. Differently, if the attribute is truly the most important one, but not the only one considered, introducing and excluding a constant in this way defined, should not distort results. Hence, in this final consistency check we make use of only a subset of the self-reported information on the importance of the attributes. What is more, model selection analysis suggests that the model making use of $ASC_{Most}$ is ought to be preferred one in this empirical application.

## 3. Survey design and data

By 2020 the European Union aims at reducing the consumption of  energy by 20%, decreasing Green House Gas emissions by 20% and increasing energy from renewable by 20% (the so-called 20-20-20 EU strategy). A great deal of initiatives is linked to this goal. One is represented by the covenant of mayors (eumayors.eu), according to which local and regional authorities define their own strategies to increase energy efficiency and the use of renewable energy sources. Once a local authority decides to be part of this initiative, a sustainable energy action plan has to be set, giving priority to the following goals: cleaner transports, requalification of public and private buildings, foster citizens' awareness regarding energy consumption.

In light of these considerations, a CE was set, aimed at assessing values citizens attach to some local public policies regarding energy efficiency, garbage collection, and transportation. Five attributes were chosen: Technology, Garbage, Emissions, Transport, and Local Tax; together with the corresponding levels, these are listed in Table 1.

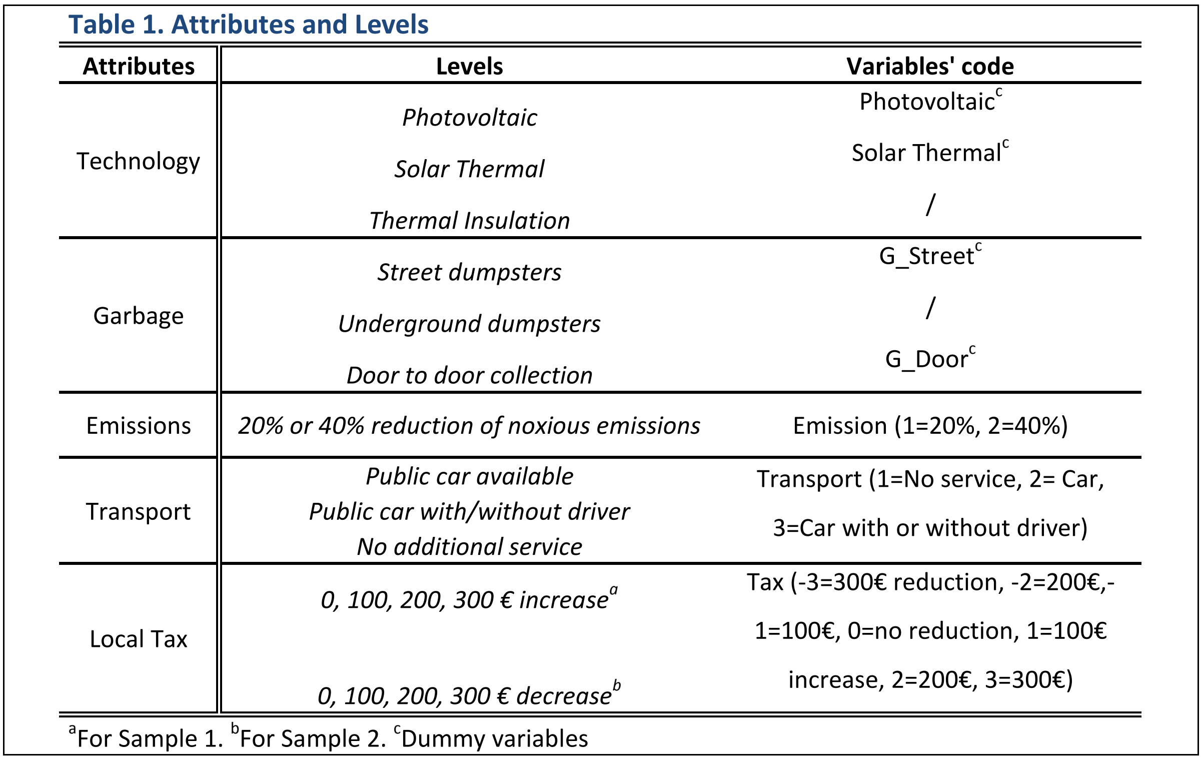


Table 1. Attributes and Levels

| Attributes | Levels | Variables' code |
|---|---|---|
| Technology | *Photovoltaic*<br>*Solar Thermal*<br>*Thermal Insulation* | Photovoltaic[c]<br>Solar Thermal[c]<br>/ |
| Garbage | *Street dumpsters*<br>*Underground dumpsters*<br>*Door to door collection* | G_Street[c]<br>/<br>G_Door[c] |
| Emissions | *20% or 40% reduction of noxious emissions* | Emission (1=20%, 2=40%) |
| Transport | *Public car available*<br>*Public car with/without driver*<br>*No additional service* | Transport (1=No service, 2= Car, 3=Car with or without driver) |
| Local Tax | *0, 100, 200, 300 € increase*[a]<br>*0, 100, 200, 300 € decrease*[b] | Tax (-3=300€ reduction, -2=200€,-1=100€, 0=no reduction, 1=100€ increase, 2=200€, 3=300€) |

[a]For Sample 1. [b]For Sample 2. [c]Dummy variables

As far as Technology and Garbage are concerned, their levels cannot be aprioristically ordered, so they will be entering the models as dummy variables. Scenarios were presented as possible outcomes of a local policy, aimed at improving energy efficiency, providing some form of garbage collection and transport service (car sharing). The policy implementation will result in emissions reduction and local tax variation.  The choice scenarios were realized by means of a $D_b$ efficient experimental design (Sàndor and Wedel 2001). Each respondent undertook six choice tasks, each with two unlabeled alternatives.

The research reported in this paper is an experimental part of the larger survey. We used a sub-sample and run a field experiment to study the follow up ranking’s usefulness in Choice

Experiments modeling[1]. The questionnaire for the field experiment has been administered face-to-face to a sample of 216 respondents in three towns of Sardinia, Italy, leading to 1296 observations. At the end of the CE, respondents have stated the order of attributes' importance. Notably, in order to make the choice task as much realistic as possible, half of the sample (S1) had increases in local tax, while the other half (S2) had tax reductions. In the estimated models a single variable that pools the levels is included as Wald test excluded the presence of significant difference with respect to increase vs decrease's coefficients.

## 4. Methodology

The data obtained from the choice experiments will be analyzed by a set of econometric models, standard in the choice modeling literature based on the Random Utility theory (McFadden 1974) and Lancaster's theory of value (Lancaster 1966). The base model, namely the Conditional Logit, assumes each respondent n has homogeneous preferences characterized by the following utility function:

$$U_{njt} = \beta_k X'_{kjt} + \varepsilon_j \tag{1}$$

where $X_{kjt}$ is the matrix of the k attributes in the choice tasks; $\beta_k$ is the vector of coefficients to be estimated and $\varepsilon_j$ is the stochastic component, independently and identically drawn from a Gumbel distribution. Given a maximizing utility behavior, the probability that individual n

[1] Table A in Appendix shows how we elicited the rankings.

[2] Therefore, we are assuming that the respondent checks in which option the best level of the most important

selects option j in choice task t is function of the deterministic component of the utility function:

$$P_{njt} = \frac{\exp\,(\beta_k X'_{kjt})}{\sum_j \exp\,(\beta_k X'_{kjt})} \tag{2}$$

Once estimated the coefficients, we will compute the willingness to pay. The ratio between the coefficient attached to a non-monetary attribute over the coefficient relative to the monetary attribute, Tax, will produce the monetary valuation, as each coefficients represent the marginal utility for the correspondent attribute:

$$WTP = -\frac{\beta_{non\,monetary}}{\beta_{Tax}} \tag{3}$$

Next, homogeneity (and independence of irrelevant alternatives' assumptions) will be dropped and checked whether the Random Parameters Logit model provides a better fit. In this framework, the researcher has to decide which parameter is deemed to be randomly distributed and according to which distribution too. This complicates the probability in (2) as we now need to integrate over all possible values of the coefficients, weighed by their density:

$$P_{njt} = \int \frac{\exp\left(\beta_{nk} X'_{kjt}\right)}{\sum_j \exp\left(\beta_{nk} X'_{kjt}\right)} f(\beta_n|\theta) d\beta_n \tag{4}$$

where $f(\beta_n|\theta)$ is the density function of the coefficients, whereas $\theta$ is a vector of parameters characterizing the deviations from the mean of the coefficients. The model is estimated by simulated maximum likelihood as the corresponding integral does not have a close form solution (McFadden and Train 2000). Conditioning upon the observed choices, it is possible to obtain, for each of the respondents, the mean coefficient and the standard deviation. We will run the model estimation assuming all the parameters to be normally distributed. Hence, conditioned on observed choices, for each attribute k and respondent n we obtain:

$$\beta_{nk} \sim N(\mu_{nk}, \sigma_{nk}^2) \tag{5}$$

Given these, for each respondent and attribute we can compute the following ratio HH (from Hess and Hensher 2010) given by:

$$HH = \frac{\sigma_{nk}}{\mu_{nk}} \tag{6}$$

High values of this ratio indicate an overly disperse individual specific normal distribution. In turn, this can suggest that the individual n might not have considered the attribute k.

Next, we estimate the Conditional and the Random Parameters Logit including the alternative specific constant $ASC_{Most}$ which is characterized in the following way:

$$\begin{cases} ASC_{Most} = 1\ if\ opt. A\ contains\ the\ best\ level\ of\ the\ most\ important\ attribute \\ or \\ ASC_{Most} = 1\ if\ opt. A\ contains\ the\ second\ best\ level\ of\ the\ most\ important\ attribute \\ ASC_{Most} = 0\ otherwise \end{cases}$$

(7)

Notably, some of the attributes have qualitative levels: technology and garbage. Suppose individual n has ranked technology as the most important attribute and photovoltaic as the most important level. Also, suppose she was presented with option A and B in choice task t without the level photovoltaic in any of the options: in this case, the constant captures the value of the second most important level of technology[2].

[2] Therefore, we are assuming that the respondent checks in which option the best level of the most important attribute is. If none of the two options in a given choice task contains that, he or she checks which one contains the second stated most important level, when it comes to the qualitative levels. For the attributes with quantitative levels, in each choice task one option contains the best level of the most important option.

Hence, the utility function in (1) becomes:

$$U_{njt} = \beta_{ASC} ASC_{Most} + \beta_k X'_{kjt} + \varepsilon_j \tag{8}$$

We also model the probability of observing a given option being chosen, in a confirmatory latent class framework. Modelling heterogeneity in a discrete framework, we estimate the probability of choosing alternative j conditional on being probabilistically assigned to segment C:

$$P_{njt|c} = \frac{\exp\,(\beta_{k|c} X'_{kjt})}{\sum_j \exp\,(\beta_{k|c} X'_{kjt})} \tag{9}$$

Constraints on the parameters will be placed to examine class with no attendance ($\beta_{k|c} = 0, \forall k \in K$), full attendance and classes in which some attributes are assumed not to have been attended.

Differently, the ranking data will be modeled by means of the ranked order Logit model. Specifically, the individual log-likelihood is given by the probability of observing a given sequence of ordered choices (see Beggs et al. 1981):

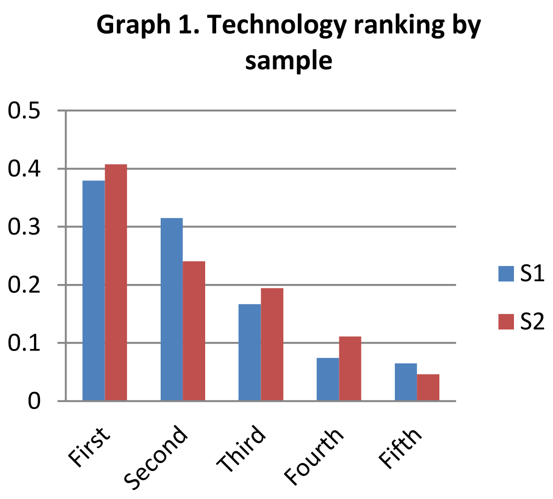
$$pr[R_1 > R_2 \ldots > R_j] = pr[R_1 > R_j\ for\ j = 2, J]pr[R_2 > R_j\ for\ j = 3, J] \ldots pr[R_{j-1} > R_j] \tag{10}$$

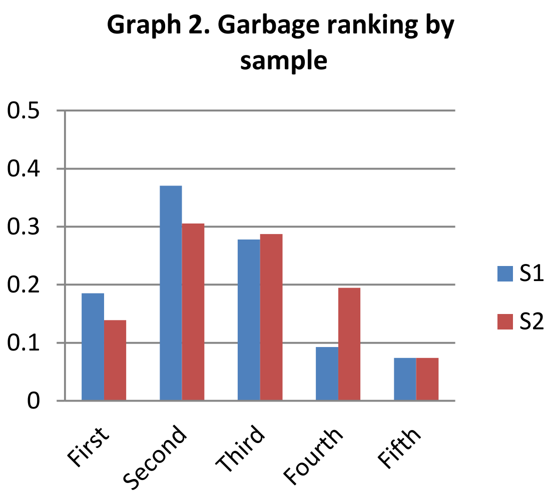
where $R_j = A_j + \varepsilon_j$. $A_j$ represents the utility attached to the attribute j while ε is the random term, following a logistic distribution.

## 4. Stated attribute importance: Results

*Descriptive statistics*

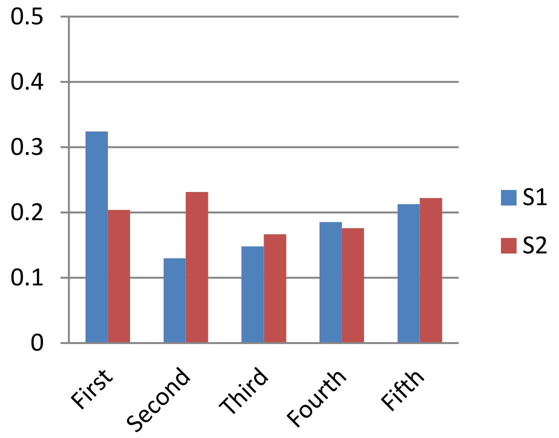
The attribute Technology is the most important, and this is true for both the two samples. Overall, almost 70% rated it as the first or the second most important (Graph 1). On the opposite end, the least important attribute is Transport (Graph 5), with almost the 70% of the respondents placing it at the last or second last position. The attribute Garbage (Graph 2) has second as the modal category for both the samples.



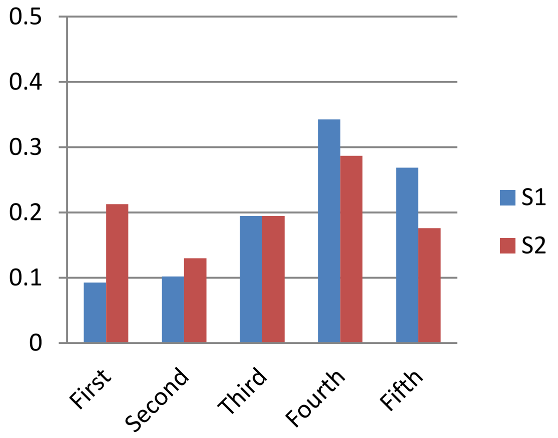


As far as the attribute Tax is concerned, 26% rated it as the most important while 22% as the least important. This attribute has been placed as first more frequently by sample 1(facing a tax increase). When it comes to Emissions (Graph 4), fourth is the modal category. Overall, the 15% ranked it as first and the 22% as the least important. Tax and Emissions are in the middle in terms of importance, compared to the other three attributes.

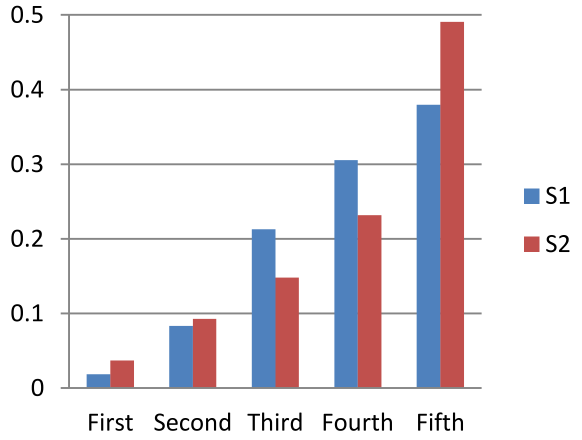
Graph 3. Tax ranking by sample



Graph 4. Emission ranking by sample

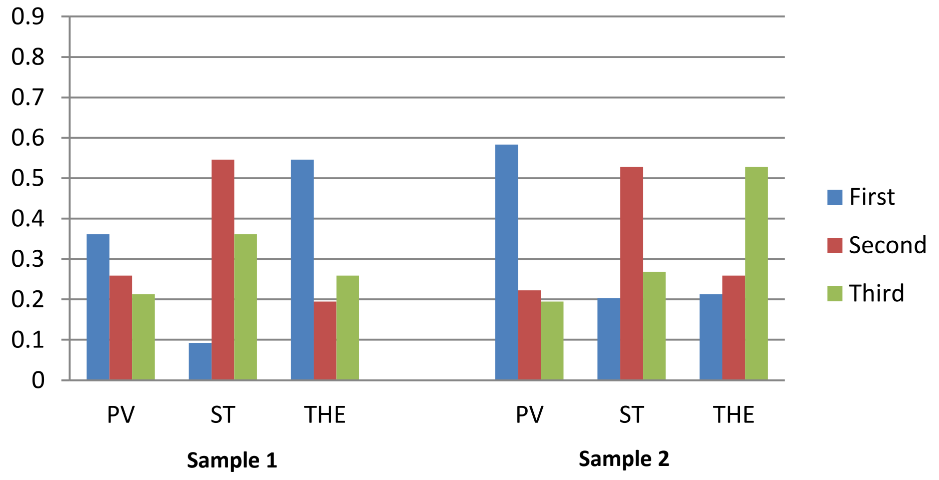


Graph 5. Transport ranking by sample

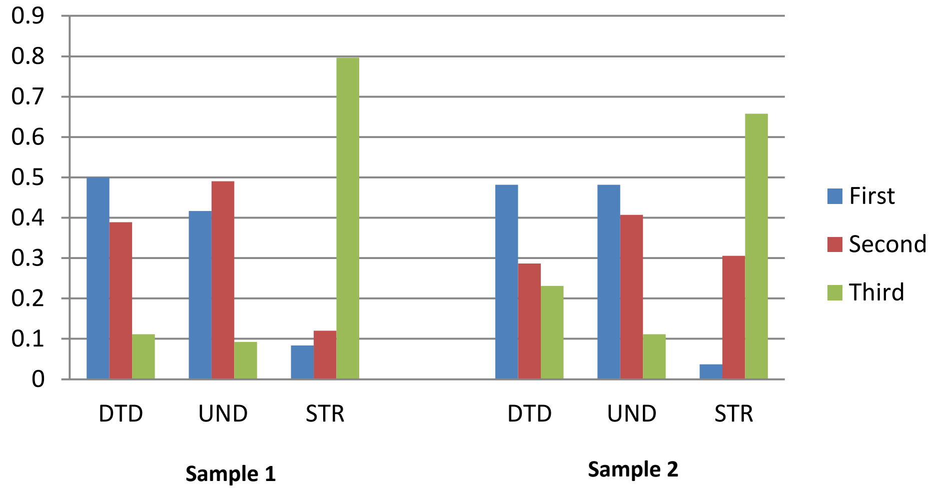


Respondents have been also asked to rank the levels of the attributes Technology and Garbage, since these are not quantitative. Starting with the former (Graph 6), Photovoltaic appears to be the most important level on average, with 47% of the sample rating it so. However, there is a clear difference between the two samples: S2 definitely prefers PV. Next, Solar Thermal is rated as second by almost half of the entire sample; Graph 6 shows how "second" is the modal category in both of the samples. Finally, Thermal Insulation is the level receiving the greatest share, about 40%, of "third". However, it is the most important technology for sample 1. Thermal insulation's measures have been undertaken by some respondents (34% on average): of these, only the 30% rated this technology as the least important level.

Graph 6. Technology sub-ranking by sample



Graph 7. Garbage sub-ranking by sample



Finally, we consider the levels of the attribute Garbage (Graph 7). There is not a conspicuous difference between Door to Door and Underground Dumpsters' rankings. In both the two samples, either first or second are the modal categories. Differently, Street Dumpsters is clearly the least important level. Summarizing, the sub-ranking of the attribute technology envisages Photovoltaic as clearly the most important one on average. Instead, for the attribute Garbage, Street Dumpsters appears to be the least important level.

*Ranked order Logit models*

We estimate three Ranked Ordered Logit (ROL) models: one (M1) for the attribute's rankings (5 levels of importance) and one each for the sub-rankings (M2 and M3), namely Technology and Garbage (3 levels each). By looking at the coefficients, shown in Table 2, we can understand whether there is a significant difference regarding the importance attached to the attributes.

In each model one attribute (or level) has to be left out as the reference one.

**Table 2. Ranked Ordered Models**

| Ranking M1 | Coeff. (Std. Err.) | Sub-Ranking (Technology) M2 | Coeff. (Std. Err.) | Sub-Ranking (Garbage) M3 | Coeff. (Std. Err.) |
|---|---|---|---|---|---|
| Technology | .796***<br>(.124) | Photovoltaic | .333***<br>(.124) | | |
| | | Thermal Ins. | .031<br>(.125) | | |
| Garbage | .410***<br>(.121) | | | Door to Door | .064<br>(.129) |
| | | | | Street Dump. | -1.52***<br>(.154) |
| Emissions | -.233***<br>(.119) | | | | |
| Transport | -.715***<br>(.126) | | | | |
| *Log Likelihood* | *-946.336* | | *-382.285* | | *-310.839* |
| *Observations* | *1080* | | *648* | | *648* |
| *N* | *216* | | *216* | | *216* |

*Tax reference category (RC). Solar Thermal RC. Underground Dumpsters RC.*
**** Statistically significant at the 1% level*

As regards M1, the attribute Tax is the reference one. Results show Technology is the most important one, followed by Garbage. Next, Emissions and Transport are less important, with the latter being the least important one.

Hence, the estimated ranking is as follows:

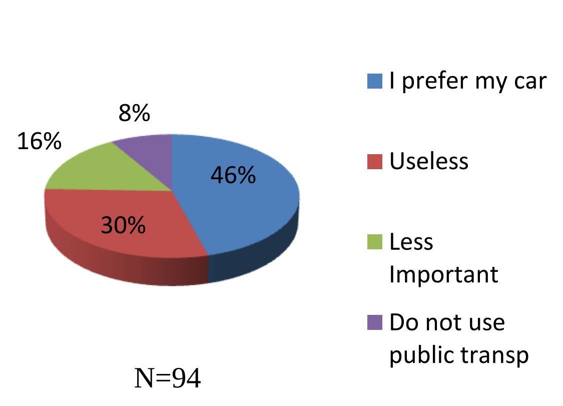
- Technology most important
- Garbage
- Tax
- Emissions
- Transport least important

This mirrors what previously observed looking at the descriptive statistics of the ranking data. As regard the sub-ranking Technology, Solar Thermal is the reference category. Results indicate there is not a significant difference between the importance attached to Solar and Thermal Insulation, while Photovoltaic is significantly more likely to be ranked first. This is also in line with the results in the previous section. Interestingly, Photovoltaic is the least common technology installed among the respondents. Finally, considering the sub-ranking Garbage, the reference category is in this case Underground Dumpsters. We previously noted that street dumpsters appeared to be the least important level; estimates from the model M3 confirm so. Furthermore, there is no significant difference between Underground and Street Dumpsters.

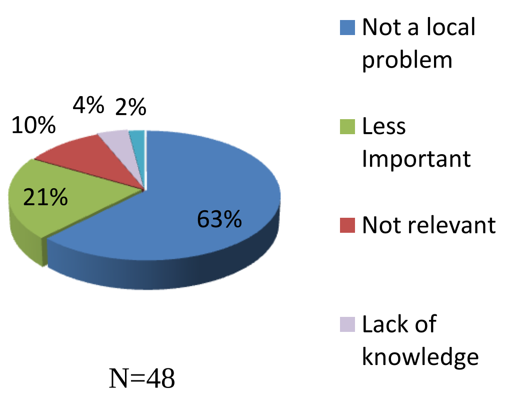
*Least important (LI) attributes: stated reasons*

We now analyze the answers given to the least important attributes, to detect whether some of the respondents have given reasons such that we might suspect they have completely ignored these attributes. Only a small percentage of the respondents literally stated that they placed that attribute as the fifth because it was the less important. But these are not the only individuals showing to have considered the attribute and attaching less importance to it. This is also true when for example we read '*I prefer my car', 'Not relevant', 'I do not use public transport'*.

**Graph 8. Stated reasons for the LI attribute: Transport**

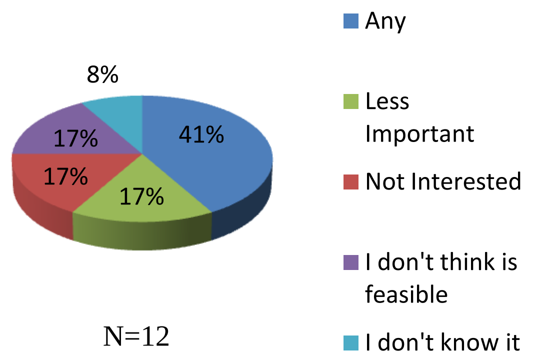


**Graph 9. Stated reasons for the LI attribute: Emissions**

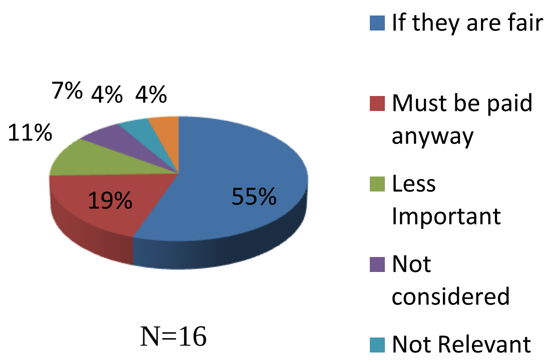


Considering the attributes Garbage and Technology, some ranked it as last simply because they are indifferent between their levels. On the other hand, it must be noticed the presence of protest attitudes. Although limited as far as Emission is concerned (*'Emissions must be reduced anyway'*), a substantial number of respondents express protest when it comes to the attribute Tax. Half of those placing it as the least important stated they would pay *'only if they are fair'*; in addition, other state that '*Taxes must be paid anyway'* and some others that they '*do not believe in tax reductions'*. Hence, although none of the respondents explicitly reported to have ignored the attribute, some have exhibited protest attitudes, therefore placing the attribute as the least important not due to their genuine preference.

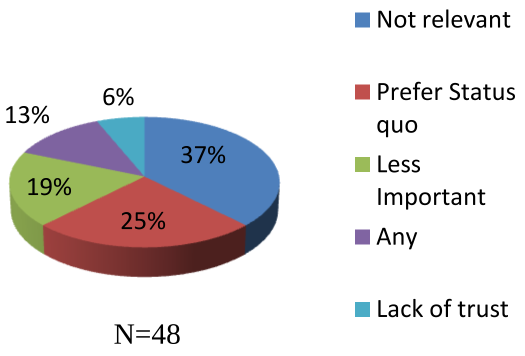
Graph 10. Stated reasons for the LI attribute: Technology



Graph 11. Stated reasons for the LI attribute: Tax



Graph 12. Stated reasons for the LI attribute: Garbage



`

## 3. Internal consistency: Results

### 3.1 Comparing WTP and explicit ranking

Table 3 shows the conditional Logit models results, along with the marginal rates of substitution and the WTP estimates, both computed with the monetary attribute, tax, as denominator.

**Table 3. Conditional Logit results**

| Variable | Coeff.<br>(Std. Err) | MRS<br>(Std. Err) | WTP<br>(€) |
|---|---|---|---|
| Photovoltaic | -.045<br>(.100) | -.168<br>(.367) | NS |
| Solar Thermal | -.151*<br>(.092) | -.558*<br>(.328) | -55.8 |
| G_Door | .018<br>(.100) | .068<br>(.372) | NS |
| G_Street | -.578***<br>(.103) | -2.13***<br>(.385) | -213 |
| Emission | .301***<br>(.064) | 1.10***<br>(.266) | 110 |
| Transport | .114**<br>(.047) | .421**<br>(.164) | 42.1 |
| Tax | -.271***<br>(.036) | / | / |
| *Log Likelihood* | *-826.42* | | |
| *Pseudo R2* | *0.08* | | |
| *Observations* | *1296* | | |
| *N* | *216* | | |

***1%, **5%, *100 significance level. NS: not significant.

We consider the CL specification as we are focusing on the average sample values, whereas heterogeneity will be taken into account in the following sections. First, we notice that the greatest (negative) WTP is in correspondence of G_street: respondents would be willing to accept 213 € to accept Street Dumpsters instead of other garbage's collection. Garbage is the second most important attribute, whereas street dumpsters was the least important level. Hence, the negative sign is in line with the explicit ranking. Furthermore, it is also confirmed there is no significant difference between door to door and underground dumpsters.

Next, considering the least important attribute, namely transport, we first note its estimated coefficient is positive and statically significant. Its associated willingness to pay is 42 euro. In absolute terms, this is the lowest WTP. Likewise, emissions' reductions are positively valued with a WTP of 110 euro to reduce emissions by 20% with respect to current levels. This amount is greater than the WTP to obtain the transport service. This is also consistent with the average ranking, as the attribute emission appears to be relatively more important than transport.

Finally, we consider the levels of the most important attribute, technology. We would have expected a positive and significant coefficient for PV, which resulted to be the most important level on average, but this is not the case: there is not a significant difference, in terms of preference, towards PV and thermal insulation. However, negative and significant is the coefficient for solar thermal: respondents would be willing to accept 55.8 euro to let the local council promote this technology as opposed to PV and insulation. Hence, overall, we find correspondence between average explicit rankings and WTP.

### 3.2 Individual parameter distributions

We now explore how much variation is present around the mean of the individual parameters. The correspondent Random Parameters Logit is shown in Table 4. All the parameters are random and assumed to be normally distributed as both negative and positive effects are plausible for each of the attributes. Mean effects are in line with the results obtained with the CL model, which assumes preference homogeneity. However, there is evidence of heterogeneity in the sample, as all standard deviations’ coefficients are statistically significant and with large magnitudes.

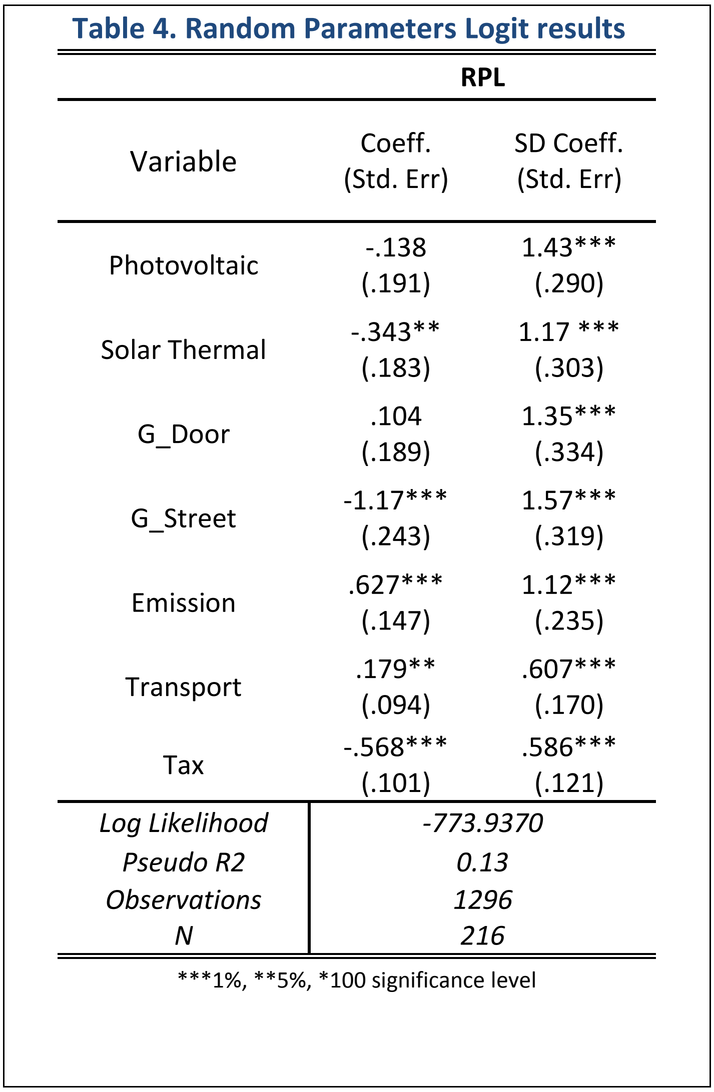


Table 4. Random Parameters Logit results

|  | RPL |  |
|---|---|---|
| Variable | Coeff.<br>(Std. Err) | SD Coeff.<br>(Std. Err) |
| Photovoltaic | -.138<br>(.191) | 1.43***<br>(.290) |
| Solar Thermal | -.343**<br>(.183) | 1.17 ***<br>(.303) |
| G_Door | .104<br>(.189) | 1.35***<br>(.334) |
| G_Street | -1.17***<br>(.243) | 1.57***<br>(.319) |
| Emission | .627***<br>(.147) | 1.12***<br>(.235) |
| Transport | .179**<br>(.094) | .607***<br>(.170) |
| Tax | -.568***<br>(.101) | .586***<br>(.121) |
| *Log Likelihood* | *-773.9370* |  |
| *Pseudo R2* | *0.13* |  |
| *Observations* | *1296* |  |
| *N* | *216* |  |

***1%, **5%, *100 significance level

Starting from this model we estimated the individual coefficients conditioned on the observed choices. Table 5 reports summary statistics for each coefficient distinguishing between those individuals who ranked it as the least important versus the rest of the sample. As warned in Scarpa et al. (2013), p-values for differences of means of conditional distributions are difficult to derive. Here, we focus on describing whether the magnitude of the means follows a given pattern depending on the stated attribute importance. Specifically, we would expect to observe a greater mean coefficient's magnitude relative to a given attribute for those respondents who ranked it as relatively more important and vice versa.

For technology and garbage’s levels, we make use of the sub-rankings, where 3 means least important and 1 most important. Instead, for Emission, Transport and Tax we look for differences between those who ranked any of these as the least important (R=5) versus the rest of the sample (R<5).

**Table 5. Sample average, standard deviation, minimum and maximum for the conditional distributions.**

| Coeff. | Ranking | N | Mean | SD | MIN | MAX |
|---|---|---|---|---|---|---|
| Photovoltaic | SR<3 | 154 | 0.037 | 0.833 | -1.91 | 1.98 |
|  | SR=3 | 62 | -0.576 | 0.771 | -2.03 | 1.74 |
| Solar Thermal | SR<3 | 148 | -0.285 | 0.634 | -1.98 | 1.64 |
|  | SR=3 | 68 | -0.508 | 0.554 | -1.81 | 0.841 |
| G_door | SR<3 | 179 | 0.202 | 0.711 | -1.49 | 1.95 |
|  | SR=3 | 37 | -0.392 | 0.637 | -1.43 | 1.46 |
| G_street | SR<3 | 59 | -0.634 | 0.803 | -2.40 | 1.71 |
|  | SR=3 | 157 | -1.39 | 0.949 | -3.33 | 1.14 |
| Emission | R<5 | 168 | 0.683 | 0.714 | -1.26 | 2.21 |
|  | R=5 | 48 | 0.413 | 0.492 | -0.864 | 1.43 |
| Transport | R<5 | 122 | 0.230 | 0.308 | -0.523 | 1.06 |
|  | R=5 | 94 | 0.119 | 0.264 | -0.617 | 0.636 |
| Tax | R<5 | 169 | -0.658 | 0.345 | -1.307 | 0.454 |
|  | R=5 | 47 | -0.328 | 0.266 | -0.905 | 0.255 |

SR=Sub-ranking. R=Ranking.

For the attributes Emission, Transport and Tax, the mean coefficients of those attaching more importance present a greater magnitude. But considering the other two statistically significant coefficients, Solar Thermal and G_street, it emerges the opposite picture: the coefficient is greater for those attaching less importance. In particular, there is a greater negative effect. The fact that a respondent places a given attribute at the end of the ranking might also hide an even stronger dislike towards that attribute as opposed to the other respondents.

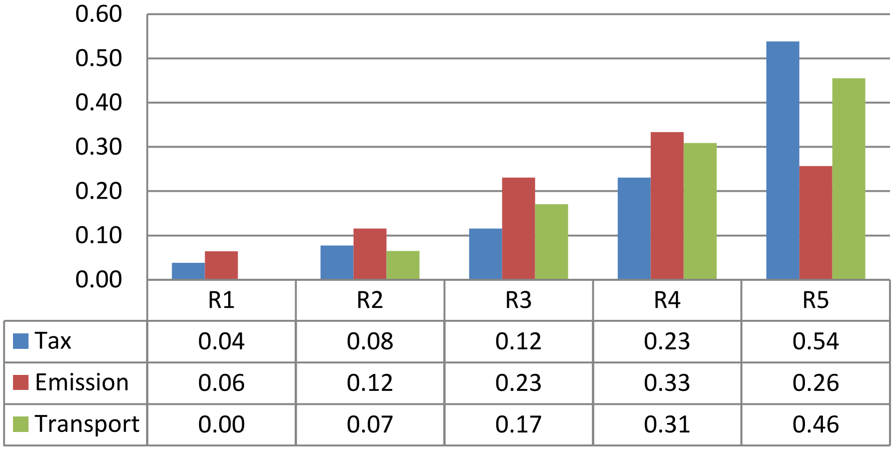
Usually, within the inferred ANA literature, conditional mean and variance are used to compute a coefficient of variation, as shown in (6), interpreted as the signal to noise ratio (Hess and Hensher 2010). Assuming the coefficients are normally distributed, which is the case in this empirical application, a value of this ratio (in absolute value) greater than 2 is interpreted as signaling attribute non-attendance as the correspondent normal distribution is over-disperse (Hess and Hensher 2010; Scarpa et al. 2013).

Graph 15 and 16 show, respectively for the attributes with quantitative and qualitative levels, the composition of individuals with a ratio greater than 2 (or lower than minus 2) for each level of importance. For each attribute, we would expect to observe a larger share of individuals for lower levels of attribute importance (where we might expect some of the individuals to have considered the attribute to a less extent while making its choice).

Considering the attributes Tax, Emissions, Transport and the Technology's level Solar Thermal, we observe an increasing trend in line with our expectations: the lower the level of importance, the more the respondents with a HH ratio greater than 2. However, for Solar Thermal the largest share is in correspondence of those who ranked this technology as the second one.

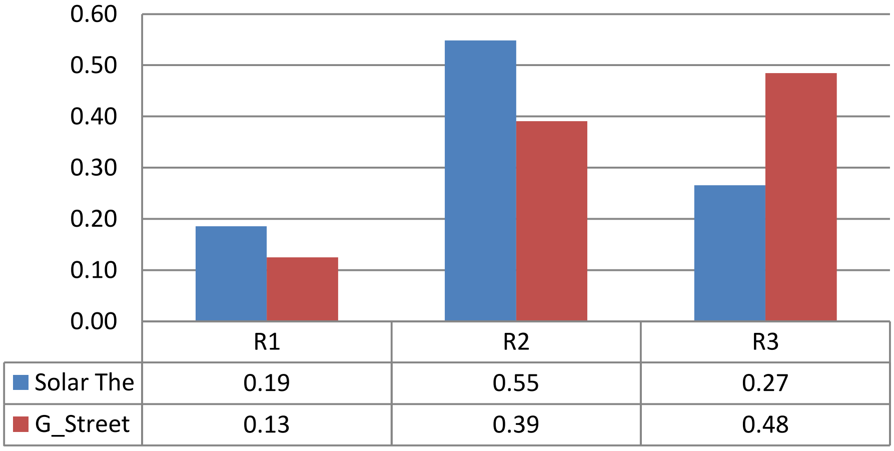
In order to determine whether there is a significant association between the rankings and the ratio we conducted a series of Pearson's chi square test and Fisher's Exact test (depending on the number of frequencies in the cells, see Appendix Tables B1-B5). Results suggest the null hypothesis of independence is rejected for all but one case: unsurprisingly, after having observed Graph 16, there is no association between the ranking of Solar Thermal and values of |HH ratio| greater than 2.

**Graph 15.|HH Ratio|>2 by Stated Attribute Importance-**
**5 levels ranking. R1=Most Important**



| | R1 | R2 | R3 | R4 | R5 |
|---|---|---|---|---|---|
| 🟦 Tax | 0.04 | 0.08 | 0.12 | 0.23 | 0.54 |
| 🟥 Emission | 0.06 | 0.12 | 0.23 | 0.33 | 0.26 |
| 🟩 Transport | 0.00 | 0.07 | 0.17 | 0.31 | 0.46 |

**Graph 16.|HH Ratio|>2 by Stated Attribute Importance-**
**3 levels ranking. R1=Most Important**



| | R1 | R2 | R3 |
|---|---|---|---|
| 🟦 Solar The | 0.19 | 0.55 | 0.27 |
| 🟥 G_Street | 0.13 | 0.39 | 0.48 |

### 3.3 RPL's heterogeneity decomposition

*First step: Principal Component Analysis*

First, we make use of principal component analysis (PCA), one applied to the attribute rankings and one applied to the sub-rankings (the levels of the qualitative attributes). PCA's results are shown in tables 6 and 7: on the basis of the variance explained, three components have been selected in both cases. Next, the score factors have been computed for each respondent.

As regards attributes' ranking, the three components have been named GA_NoTax, Tra_NoEm and Tech_NoEm. This is because the first component has a high positive correlation with the Garbage ranking and a high negative one with the Tax ranking; the second component is characterized by a high positive correlation with the Transport ranking and a high negative one for the Emission ranking. Finally, the third one has a high positive coefficient with respect to the Technology ranking and a negative one for the Emission ranking, although lower than the analogous one for the second component.

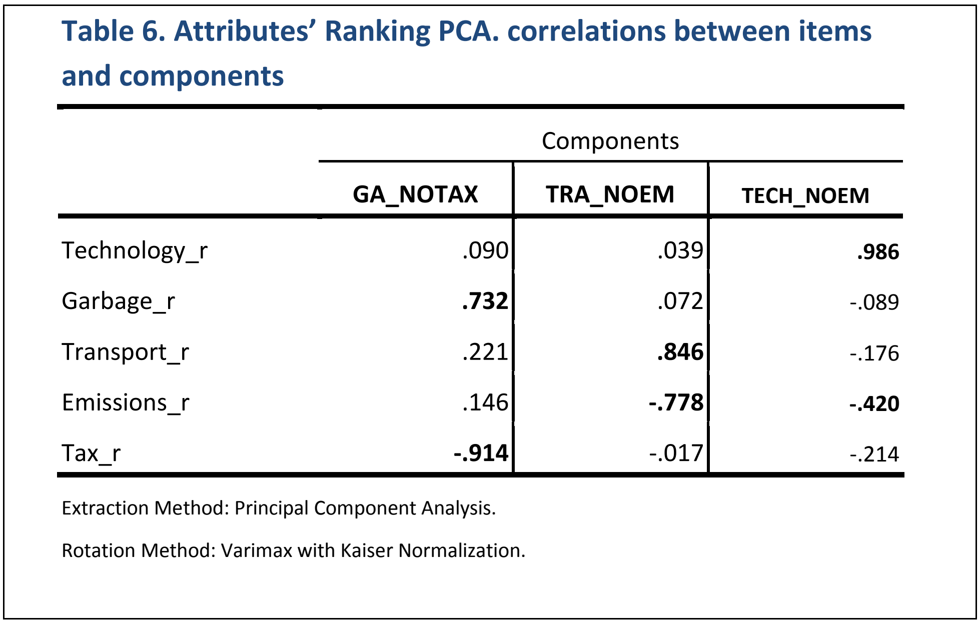


Table 6. Attributes' Ranking PCA. correlations between items and components

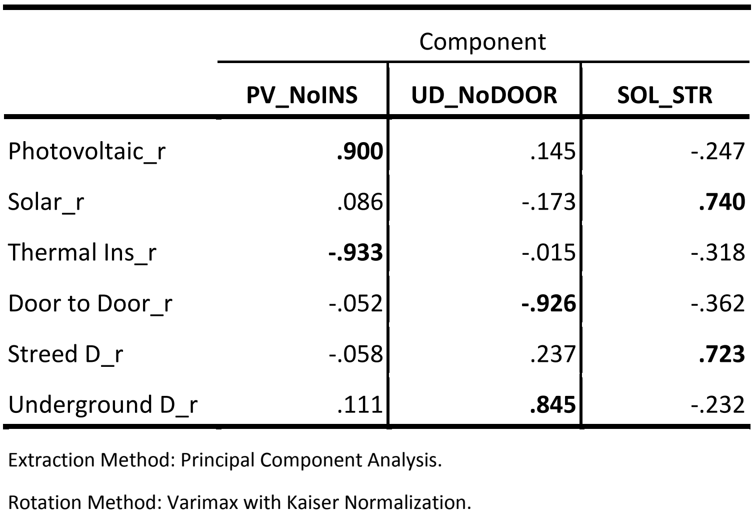
| | Components | | |
|---|---|---|---|
| | GA_NOTAX | TRA_NOEM | TECH_NOEM |
| Technology_r | .090 | .039 | **.986** |
| Garbage_r | **.732** | .072 | -.089 |
| Transport_r | .221 | **.846** | -.176 |
| Emissions_r | .146 | **-.778** | **-.420** |
| Tax_r | **-.914** | -.017 | -.214 |

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

.

**Table 7. Attributes' Sub-Ranking PCA: correlations between items and components**



| | Component | | |
|---|---|---|---|
| | PV_NoINS | UD_NoDOOR | SOL_STR |
| Photovoltaic_r | **.900** | .145 | -.247 |
| Solar_r | .086 | -.173 | **.740** |
| Thermal Ins_r | **-.933** | -.015 | -.318 |
| Door to Door_r | -.052 | **-.926** | -.362 |
| Streed D_r | -.058 | .237 | **.723** |
| Underground D_r | .111 | **.845** | -.232 |

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

Next, the PCA applied to the Sub-ranking data lead to the three following components (Table 7): PV_NoIns, UD_NoDoor and Sol_Str. The first component has a high positive correlation with Photovoltaic and high negative correlation with Thermal Insulation; differently, the second one has a high positive coefficient with respect to Underground Dumpsters and really low coefficient for the Door to door ranking. Finally, the third component has high and positive correlation coefficients for Solar Thermal and Street Dumpsters' rankings.

*Second step: Combining CE data and score factors*

We used two strategies to combine CE data and these components obtained from the rankings: a LC model in which the membership class probability is function of the components and a RPL model where the components are used to reveal the preference heterogeneity around the mean. In both cases the best model specification was obtained excluding one component from each of the two PCAs, specifically dropping Tech_NoEmi and Sol_Str. These are the "third" components in both the two PCA, hence explaining the lowest amount of variance in the data. The best RPL specification found has a Log-Likelihood of -678.6728 with 24 parameters, compared to the best LC specification (3 classes) with a value of -736.3475 with ten additional parameters. The statistical criteria AIC, AIC3, BIC and CAIC pointed towards the RPL. The same indication, at a significance level α=1%, is provided by the Ben-Akiva and Swait test (Ben Akiva and Swait 1986) for strictly non-nested models. Since the criteria point towards the same model we can be quite confident in selecting the RPL model (Strazzera et al. 2013). It is therefore here presented the RPL specification only (Table 8). As in the previous RPL model, all parameters are assumed to be normally distributed.

The interactions are all significant and all the signs are in the expected direction. When it comes to the Technology levels, these have been interacted with the component PV_NoIns. Individuals who are more likely to rank Photovoltaic as more important and Thermal Insulation as less important are associated with a positive effect on the mean of the variable Photovoltaic and Solar Thermal.

**Table 8. Random Parameters Logit with preference heterogeneity decomposition**

| Variable | Coeff.<br>(Std. Err) | SD<br>(Std. Err) |
|---|---|---|
| Photovoltaic | -.066<br>(.155) | 1.05***<br>(.257) |
| Solar Thermal | -.269*<br>(.158) | 1.00***<br>(.293) |
| G_Door | .100<br>(.154) | .589**<br>(.286) |
| G_Street | -.958***<br>(.183) | 1.15***<br>(.255) |
| Emission | .555***<br>(.124) | .949***<br>(.185) |
| Transport | .181**<br>(.082) | .504***<br>(.149) |
| Tax | -.509***<br>(.071) | .362***<br>(.104) |
| Photovoltaic*PV_NoIns | .962***<br>(.176) | / |
| Solar Thermal*PV_NoIns | .456***<br>(.155) | / |
| G_Door*UD_NoDoor | -.951***<br>(.158) | / |
| G_Street*GA_NoTax | -.391***<br>(.152) | / |
| Emission*TRA_NoEmi | -.459***<br>(.123) | / |
| Transport*TRA_NoEmi | .342***<br>(.084) | / |
| Tax*GA_NoTax | .399***<br>(.066) | / |
| *Log-Likelihood* | -687.1037 | |
| *Pseudo R2* | 0.235 | |
| *Observations* | 1296 | |
| *N* | 216 | |

***1%, **5%, *10% significance level

Next, considering Garbage's levels, G_Door has been interacted with UD_NoDoor while G_Street with GA_NoTax. Respondents more likely to have given more importance to Underground Dumpsters and less to Door to Door collection are associated with a negative effect on the mean of G_Door. Those who (more likely) ranked the attribute Garbage as more important have a negative effect on the mean of G_Street; this further increases the negative magnitude of this coefficient.

As regards the attribute Emission, a significant and negative effect has arisen from the interaction with TRA_NoEm, lowering the magnitude of the correspondent coefficient for those more likely to have given less important to the attribute Emission. As for Transport, positive is the effect attached to the interaction with TRA_NoEm, increasing the magnitude of the correspondent coefficient for those respondents who ranked Transport as more important. Finally, a positive value is associated to the mean of the attribute Tax when interacted with GA_NoTax, reducing the magnitude of the corresponding Tax coefficient for those who are more likely to have attached less importance to this attribute. These results show that heterogeneity in CE data is successfully explained by means of the associated ranking data. Remarkably, the results obtained in a latent class setting are also in support of this finding.

### 3.4 Constrained Latent Class Model

Attribute non-attendance has been often modeled by means of constrained latent class models. This modeling strategy belongs to the so-called inferred approach, namely avoiding the use of ANA statements but rather infer choice behavior by means of a suitable econometric specification. We here adopt this approach, compute respondent-specific membership probabilities and analyze the composition of each class in terms of stated attribute importance.

A given coefficient is either constrained to be equal to zero or to take the same value in all the classes where the coefficient is different from zero (Scarpa et al. 2013).

In choosing the best specification we follow Lagarde (2013), dropping those class' specification representing specific ANA combinations with zero probability. The best specification found is presented in Appendix, Table C. The confirmatory approach lead to a 4-classes model, characterized in the following way:  the first class (Class 1) is the full attendance one; Class 2 has emission, transport and tax coefficients constrained to zero; Class 3 has Garbage's levels constrained to zero and, finally, Class 4 assumes full non-attendance, i.e. all coefficients are constrained to zero. Differently from the previous literature, we have now the opportunity to check the composition of each of these classes in terms of stated attribute importance. Our expectations are the following: Class 4 should be characterized by a greater share of "5" and "4" (last positions in the ranking) compared to those observed in the other classes; conversely, Class 1 should include with a greater probability those individuals who ranked as relatively more important the attributes, as opposed to the other classes; considering class 3, respondents more likely to belong to these segment are expected to attach less importance to the attributes emission, transport and tax; finally, class 3 should include respondents who have attached less importance to the garbage's levels.

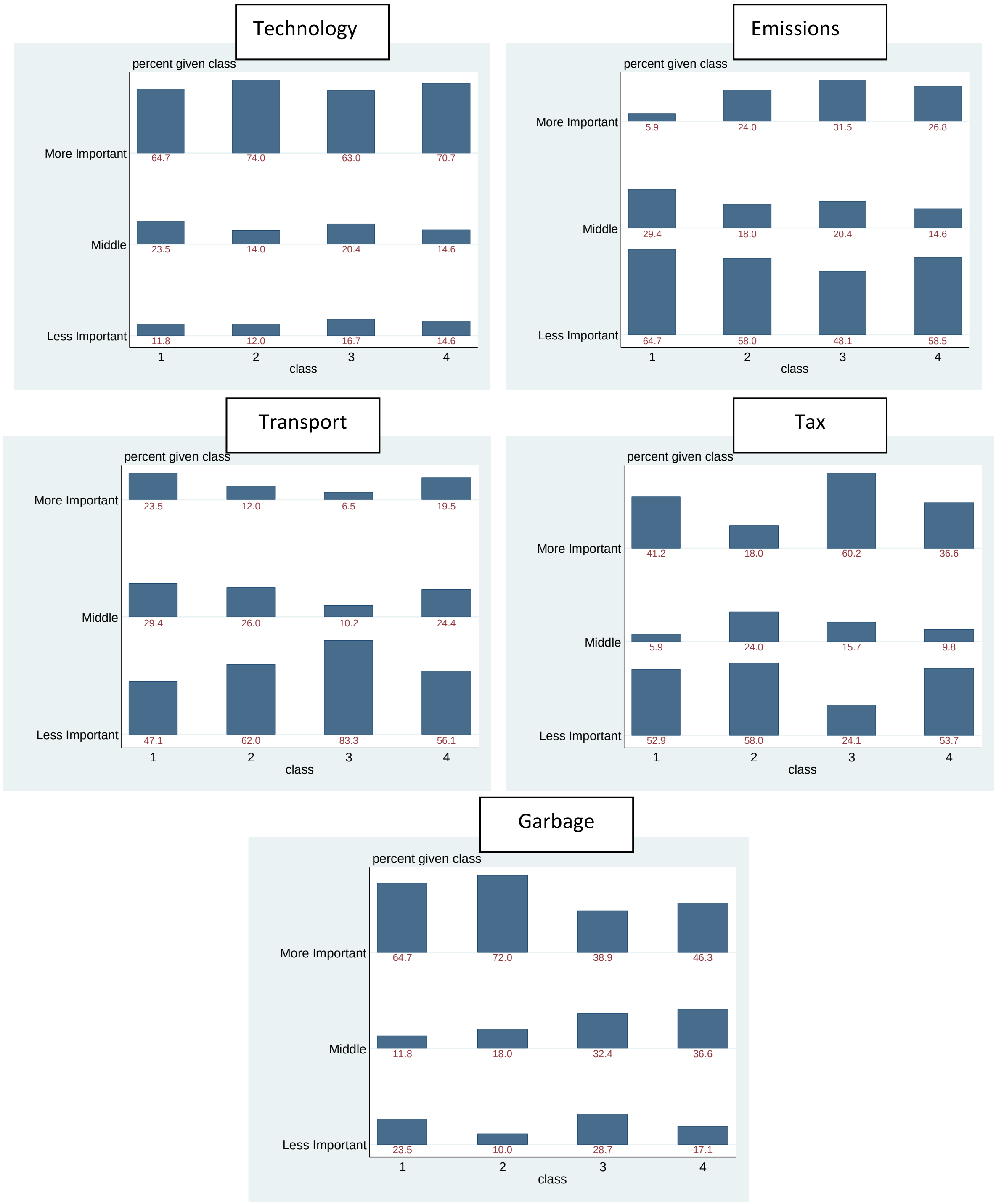
In order to assess whether there is correspondence between the classes found and the SAI, we first check whether there is a significant association between class composition and stated ranking. We run the test considering three categories: attributes ranked as first or second (More Important); attributes ranked as third (Middle) and, finally, attributes ranked as fourth and fifth (Less Important). Results of Fisher Exact test indicate that there is a significant association between class composition and Tax, Transport and Garbage rankings. Differently, there is no significant association between class composition and Technology or Emission rankings.

Subsequently, we take a closer look at the classes' composition, which are show in the Graph 17. As regards Technology and Emission's stated importance, class composition appears to be fairly constant. As far as the other three attributes are concerned, we have found significant association. Considering Transport, we would expect to observe a greater share of 'less important' in class 2 and 4 as opposed to the full attendance class 1: this is the case, although the largest share is observed in class 3, where garbage's level are constrained to zero. Still on the garbage's levels, the share of those belonging to the category 'middle' is larger in the classes 2 and 4, whereas it is lower the share of those in the category 'more important'.

Finally, considering the attribute Tax, there is a substantial difference between class 1, 2 and 4 as opposed to class 3. The latter is characterized by the greatest share of those ranking Tax as more important. Class 2 and 4, in which the respective coefficient is constrained to zero, have a greater share of 'less important'.

All in all, there is some degree of correspondence between class composition and SAI. However, this inferred ANA approach does not appear to be satisfactory in this empirical application. Across all the attributes, a substantial share of the respondents belonging to the full non-attendance class has ranked them as more important. Hence, a substantial number of respondents who have indicated a given attribute as fairly important to them might be nonetheless probabilistically associated to a class where the respective coefficients are instead constrained to zero.

**Graph 17: Classes composition by SAI**



### 3.5 An alternative specific constant to identify the most important attribute

We now focus on the attributes ranked as the most important ones. We created a constant ($ASC_{Most}$) which equals 1 if option A contains the better level of the most important attribute compared to Option B, 0 otherwise. This is straightforward for the attributes Emissions, Transport and Tax. However, when it comes to Technology and Garbage, we need to take into account the sub-rankings. Notably, there are situations in which none of the two options contain the most important Technology or Garbage's level. Hence, in these cases the constant takes the value 1 if Option A contains the second most important Technology or Garbage level. Table 9 shows the results obtained estimating a CL and a RPL including the constant $ASC_{Most}$.

We found an impressive improvement in the Log-Likelihoods of the models: compared to the models without the constant, CL's and RPL's have both about a 16% improvement with one extra parameter. Nevertheless, coefficients' magnitude and significance is not particularly affected: this is against the presence of lexicographic behavior. As far as $ASC_{Most}$ 's coefficient is concerned, its sign is in the expected direction: the alternative containing the most important (or relative more important) attribute is more likely to be chosen. In addition, the inclusion of this constant improves models' prediction: correctly predicted choices reach the 66.5% compared to almost the 57% without the constant (for both the CL and RPL model).

This, together with the noteworthy improvement in the goodness of fit and sign of the constant's coefficient, shows the presence of consistency between respondents' preferences elicited during the choice experiment and the most important attributed indicated in the ranking exercise.

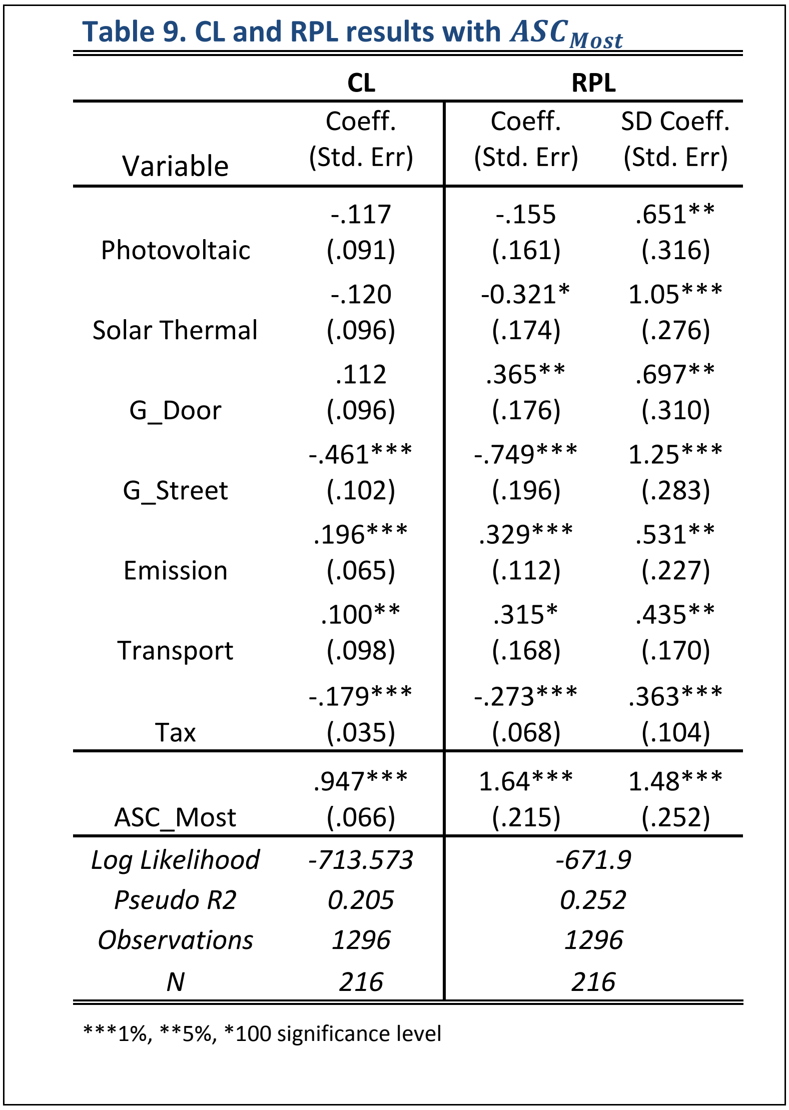


Table 9. CL and RPL results with $ASC_{Most}$

| | CL | RPL | |
|---|---|---|---|
| Variable | Coeff.<br>(Std. Err) | Coeff.<br>(Std. Err) | SD Coeff.<br>(Std. Err) |
| Photovoltaic | -.117<br>(.091) | -.155<br>(.161) | .651**<br>(.316) |
| Solar Thermal | -.120<br>(.096) | -0.321*<br>(.174) | 1.05***<br>(.276) |
| G_Door | .112<br>(.096) | .365**<br>(.176) | .697**<br>(.310) |
| G_Street | -.461***<br>(.102) | -.749***<br>(.196) | 1.25***<br>(.283) |
| Emission | .196***<br>(.065) | .329***<br>(.112) | .531**<br>(.227) |
| Transport | .100**<br>(.098) | .315*<br>(.168) | .435**<br>(.170) |
| Tax | -.179***<br>(.035) | -.273***<br>(.068) | .363***<br>(.104) |
| ASC_Most | .947***<br>(.066) | 1.64***<br>(.215) | 1.48***<br>(.252) |
| *Log Likelihood* | *-713.573* | *-671.9* | |
| *Pseudo R2* | *0.205* | *0.252* | |
| *Observations* | *1296* | *1296* | |
| *N* | *216* | *216* | |

***1%, **5%, *100 significance level

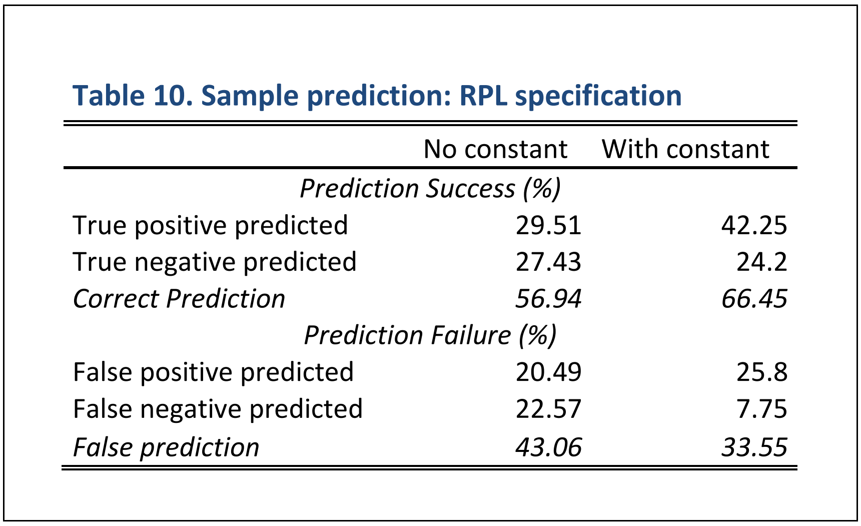


Table 10. Sample prediction: RPL specification

| | No constant | With constant |
|---|---|---|
| *Prediction Success (%)* | | |
| True positive predicted | 29.51 | 42.25 |
| True negative predicted | 27.43 | 24.2 |
| *Correct Prediction* | *56.94* | *66.45* |
| *Prediction Failure (%)* | | |
| False positive predicted | 20.49 | 25.8 |
| False negative predicted | 22.57 | 7.75 |
| *False prediction* | *43.06* | *33.55* |

These results are also quite interesting as far as we consider the decision process of the respondents in ranking the attributes. It seems that the information related to the most important attribute is more reliable than the full ranking. This may be due to the fact that 1) it is easier to recall what attribute is the most important and 2) their choice was guided by the presence of the better level of the most important attribute.

## 7. Conclusions

The results obtained from the internal consistency checks suggest there is correspondence between what the respondents choose whilst selecting the alternative during the choice experiment and how, afterwards, they rank the attributes according to the level of importance. Specifically, we found consistency between mean WTP and average explicit rankings. Considering instead the conditional distribution after having allowed for heterogeneity in a RPL framework, results show that for some attributes the greater the importance, the greater the coefficient's magnitude. But when the coefficient is negative, its magnitude gets amplified for those respondents ranking the relevant attribute as less important. This raises questions on the ranking process, as it appears that some individuals ranking a given attribute as the least important are not just ignoring it, but they might dislike it a lot: hence, having an important role in their choice indeed.

Complementing ranking with the CE data may lead to noteworthy improvements in both model fit and sample prediction. Specifically, including an alternative specific constant indicating which of the alternatives contains the better level of the most important attribute, has proven to be the best modeling strategy in this application. It has a better fit and less parameters compared to the RPL with heterogeneity decomposition through the covariates obtained applying PCA to the ranking. In addition, the RPL with the inclusion of $ASC_{Most}$ shows a notable improvement in terms of sample prediction.  Model selection criteria AIC, AIC3, CAIC and BIC, as well as the Ben-Akiva and Swait test confirm the RPL model with the alternative specific constant represents the best specification (the RPL with heterogeneity decomposition was found to be preferred to the best LC specification by means of the same criteria).

Whilst making use of different modeling strategies, we have estimated a Constrained Latent Class model, often employed within the inferred ANA literature. When SAI is to be considered a reliable self-reported information, results suggest the researcher should be cautious as the non-attendance class may actually (probabilistically) contain individuals attaching importance to the attributes.

More work is under way to confirm the reliability of this self-reported information. Further research should be aimed at testing whether checking for consistency between CE data and self-reported information could also represent a powerful tool at the pilot/pre-testing stage, in order to detect anomalies in the survey, such as an excessively complex choice task. Another line of research we envisage is to understand which variables drive the ranking, using structural equations that relate rankings with socio-economic and attitudinal variables. Hence, the next step could be represented by integrating CE and ranking data within a hybrid modeling approach. In addition, more empirical applications are needed to assess the use and validity of $ASC_{Most}$.

## Appendix

### Table A. Ranking Exercise.

Please, order the following attributes on the basis of how much you have taken them into consideration whilst making your choices.

*1=Most Important* *1=Most Important*

*5= Least Important* *3=Least Important*

□ Photovoltaic

□ Technology □ Solar Thermal

□ Thermal Insulation

□ Door to door

□ Garbage □ Street Dumpsters

□ Underground Dumpsters

□ Transports

□ Emissions

□ Tax

***Please state the reason of the first position***

***Please state the reason of the last position***

### Table B1. Chi square test-Emissions

| | Emission R | | | | |
|---|---|---|---|---|---|
| | R1 | R2 | R3 | R4 | R5 |
| **\|HH ratio\|<2** | 28 | 16 | 24 | 42 | 28 |
| **\|HH ratio\|>2** | 5 | 9 | 18 | 26 | 20 |

Pearson chi2 (4)= 7.88 Pvalue=.096

### Table B2. Fisher's Exact test-Transport

| | Transport R | | | | |
|---|---|---|---|---|---|
| | R1 | R2 | R3 | R4 | R5 |
| **\|HH ratio\|<2** | 6 | 9 | 20 | 20 | 38 |
| **\|HH ratio\|>2** | 0 | 10 | 19 | 38 | 56 |

Fisher's Exact=0.022

### Table B3. Fisher's Exact test-Tax

| | Tax R | | | | |
|---|---|---|---|---|---|
| | R1 | R2 | R3 | R4 | R5 |
| **\|HH ratio\|<2** | 56 | 37 | 30 | 32 | 30 |
| **\|HH ratio\|>2** | 1 | 2 | 4 | 7 | 17 |

Fisher's Exact=0.000

### Table B4. Chi square test-Emissions

| | Street R | | |
|---|---|---|---|
| | R1 | R2 | R3 |
| **\|HH ratio\|<2** | 5 | 20 | 126 |
| **\|HH ratio\|>2** | 8 | 26 | 31 |

Pearson chi2 (2)= 29.37  Pvalue=.000

### Table B5. Chi square test-Solar Thermal

| | Solar Thermal R | | |
|---|---|---|---|
| | R1 | R2 | R3 |
| **\|HH ratio\|<2** | 11 | 54 | 38 |
| **\|HH ratio\|>2** | 21 | 62 | 30 |

Pearson chi2 (2)= 4.16 Pvalue=.125

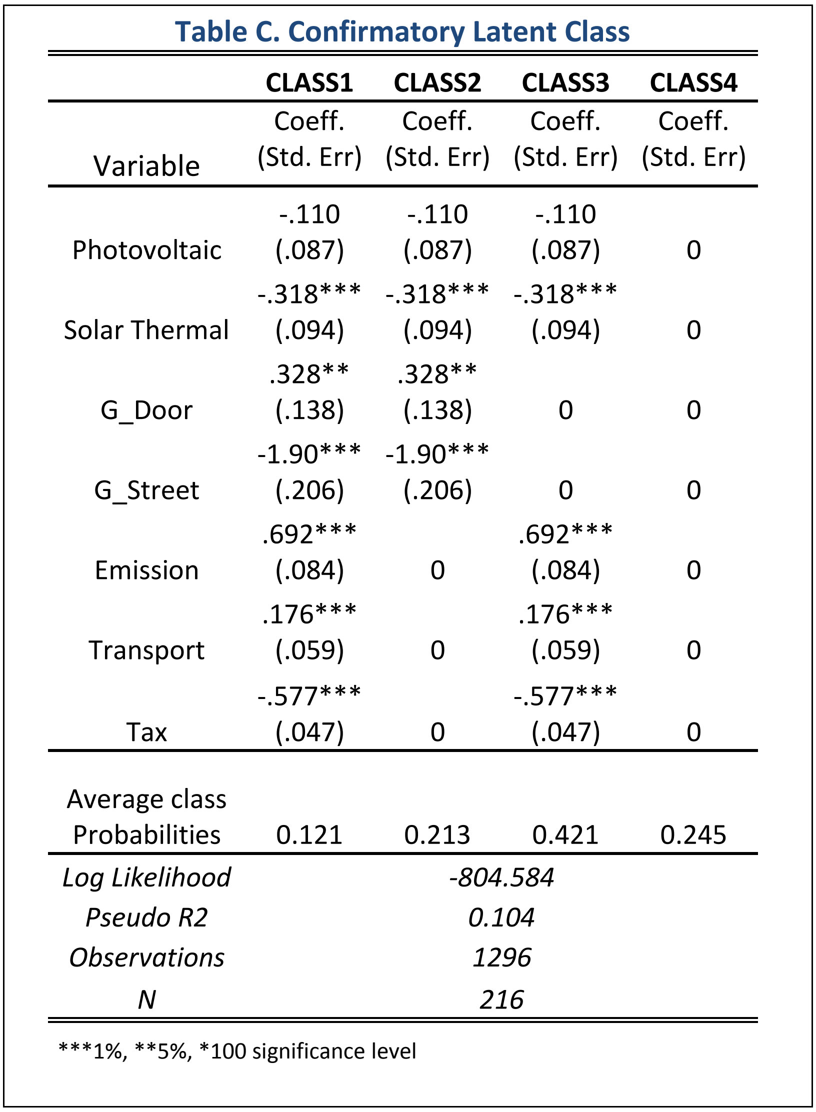


Table C. Confirmatory Latent Class

| | CLASS1 | CLASS2 | CLASS3 | CLASS4 |
|---|---|---|---|---|
| Variable | Coeff.<br>(Std. Err) | Coeff.<br>(Std. Err) | Coeff.<br>(Std. Err) | Coeff.<br>(Std. Err) |
| Photovoltaic | -.110<br>(.087) | -.110<br>(.087) | -.110<br>(.087) | 0 |
| Solar Thermal | -.318***<br>(.094) | -.318***<br>(.094) | -.318***<br>(.094) | 0 |
| G_Door | .328**<br>(.138) | .328**<br>(.138) | 0 | 0 |
| G_Street | -1.90***<br>(.206) | -1.90***<br>(.206) | 0 | 0 |
| Emission | .692***<br>(.084) | 0 | .692***<br>(.084) | 0 |
| Transport | .176***<br>(.059) | 0 | .176***<br>(.059) | 0 |
| Tax | -.577***<br>(.047) | 0 | -.577***<br>(.047) | 0 |
| Average class Probabilities | 0.121 | 0.213 | 0.421 | 0.245 |
| *Log Likelihood* | *-804.584* | | | |
| *Pseudo R2* | *0.104* | | | |
| *Observations* | *1296* | | | |
| *N* | *216* | | | |

***1%, **5%, *100 significance level

## References

Alemu MH, Mørkbak MR, Olsen SB, Jensen CL (2013) Attending to the Reasons for Attribute Non-attendance in Choice Experiments. Environ Resour Econ 54:333–359

Azevedo C, Corrigan J R, Crooker J (2009) Testing for Internal Consistency of Choice Experiments Using Explicit Rankings of Quality Attributes. In Edelstein A, Bär D Handbook of Environmental Research 507-517. Nova Science Publisher New York

Balcombe K, Bitzios M, Fraser I (2012). Using attribute importance rankings within discrete choice experiments: an application to valuing bread attributes. 86th Annual Conference of the Agricultural Economics Society, Warwick

Balcombe K, Burton M, Rigby D (2011) Skew and attribute non-attendance within the Bayesian mixed Logit model. J Environ Econ Manage 62:446–461

Beggs S, Cardell S, Hausman JA (1981) Assessing the potential demand for electric cars. Journal of Econometrics 17:1-19

Ben-Akiva M, Swait J (1986) The Akaike likelihood index. Transportation Science 20:133-136

Cameron TA, DeShazo JR (2011) Differential Attention to Attributes in Utility-Theoretic Choice Models. J Choice Model 3:73–115

Campbell D, Hensher D a., Scarpa R (2011) Non-attendance to attributes in environmental choice analysis: a latent class specification. J Environ Plan Manag 54:1061–1076

Campbell D, Lorimer VS (2009) Accommodating attribute processing strategies in stated choice analysis : do respondents do what they say they do? 17th Annual Conference of the European Association of Environmental and Resource Economics, Amsterdam

Carlsson F, Kataria M, Lampi E (2010) Dealing with Ignored Attributes in Choice Experiments on Valuation of Sweden’s Environmental Quality Objectives. Environ Resour Econ 47:65–89

Colombo S, Christie M, Hanley N (2013) What are the consequences of ignoring attributes in choice experiments? Implications for ecosystem service valuation. Ecol Econ 96:25–35

Contu D, Strazzera E (2014) Follow up statements in choice modeling: Assessment of reliability and use in model estimation. World Congress of Environmental and Resource Economists-Submitted.

Flynn T N (2010). Using conjoint analysis and choice experiments to estimate QALY values: issues to consider. Pharmacoeconomics 28:711-722

Hensher D A, Greene W (2003) The mixed Logit model: the state of practice. Transportation 30:130-176

Hensher DA, Greene WH (2010) Non-attendance and dual processing of common-metric attributes in choice analysis: a latent class specification. Empir Econ 39:413–426

Hensher DA, Rose JM, Greene WH (2005) The implications on willingness to pay of respondents ignoring specific attributes. Transp J 32:203–222

Hensher DA, Rose JM, Greene WH (2012) Inferring attribute non-attendance from stated choice data: implications for willingness to pay estimates and a warning for stated choice experiment design. Transportation 39:235–245

Hess S (2012) Impact of unimportant attribute in stated choice surveys. Transport Research Board, 91st Annual Meeting Washington

Hess S, Hensher DA (2010) Using conditioning on observed choices to retrieve individual-specific attribute processing strategies. Transp Res Part B Methodol 44:781–790

Hess S, Rose JM (2007). A latent class approach to recognizing respondents’ information processing strategies in SP studies. Oslo Workshop on valuation methods in transport planning, Oslo 2007

Hess S, Stathopoulos A, Campbell D, et al. (2013) It’s not that I don’t care, I just don’t care very much: confounding between attribute non-attendance and taste heterogeneity. Transportation 40:583–607

Hole AR, Kolstad JR, Gyrd-Hansen D (2013) Inferred vs. stated attribute non-attendance in choice experiments: A study of doctors’ prescription behaviour. J Econ Behav Organ 96:21–31

Kehlbacher A, Balcombe K, Bennett R (2013) Stated Attribute Non-attendance in Successive Choice Experiments. J Agric Econ 64:693–706

Kragt ME (2013) Stated and Inferred Attribute Attendance Models: A Comparison with Environmental Choice Experiments. J Agric Econ 64:719–736

Lagarde M (2013) Investigating attribute non-attendance and its consequences in choice experiments with latent class models. Health Econ 22:554-567

Lancaster K (1966) A New Approach to Consumer Theory. J Polit Econ 74:132–157

McFadden D (1974) Conditional Logit analysis of qualitative choice behavior. In Zarembka P-Frontiers in Econometrics. Academic Press: New York

McFadden D, Train K (2000) Mixed MNL models for discrete responses. Journal of Applied Econometrics 15:447-470

Sándor Z, Wedel M (2001) Designing conjoint choice experiments using managers’ prior beliefs. Journal of Marketing Research 38:430-444

Sælensminde K (2001) Inconsistent choices in stated choice data: use of the Logit scaling approach to handle resulting variance increases. Transportation 28:269-296

Scarpa R, Gilbride TJ, Campbell D, Hensher DA (2009) Modelling attribute non-attendance in choice experiments for rural landscape valuation. Eur Rev Agric Econ 36:151–174

Scarpa R,  Zanoli R, Bruschi V, Naspetti S (2013) Inferred and Stated attribute non-attendance in food choice experiments. American Journal of Agricultural Economics 95:165-180

Strazzera E, Contu D, Ferrini S (2013) Check it out! A Monte Carlo analysis of the performance of selection criteria and tests for choice experiments models. International Choice Modeling Conference, Sydney 2013

Swait J (1994) A structural equation model of latent segmentation and product choice for cross-sectional revealed preference choice data. Journal of Retail and Consumer Services 1:77-89

# **Chapter 3:** Task complexity in choice experiments: A review

## Task complexity in choice experiments: A review[*]

### Abstract

Whilst designing a discrete choice experiment, every practitioner has to set the complexity of the task the respondents are going to undertake. Many studies have dealt with this issue, but most of them have focused only on partial definitions of task complexity. Nevertheless, some studies have gone as far as proposing some numbers (e.g.: the number of choice tasks and the attributes) which can be considered feasible or "best practice". Here, we focus on the structuralist view of Task Complexity in choice experiments. After having provided a systematic review of the literature we find that the number of the attributes, the number of attributes' levels and the number of alternatives are reported to significantly affect error variance and willingness to pay, whereas less substantial impacts emerge with respect to the number of choice tasks. Nevertheless, results are country specific. We suggest it is quite risky to define what a high number is ex-ante: this has to be determined in the building of the DCE. Finally, considering 30 articles recently published in the Transportation, Environmental and Health Economics literature, we find that although the issue of task complexity is acknowledged, it is not treated satisfactorily.

**Keywords**: Task Complexity, Discrete Choice Experiments, Error Variance.

### Abbreviations

| | |
|---|---|
| ADM_M | Administration mode |

[*] I am especially grateful to John Rose for the help received whilst visiting the Institute of Transport and Logistics Studies, University of Sydney Business School. The presence of any error is my sole responsibility.

| | |
|---|---|
| ATT_R | Attributes’ level range |
| AVCM | Asymptotic Variance Covariance Matrix |
| CA | Constant Alternative |
| DCE | Discrete Choice Experiment |
| ED | Experimental Design |
| ENV | Environmental Economics |
| GMNL | Generalized Multinomial Logit |
| HE | Health Economics |
| H-MNL | Heterosckedastic Multinomial Logit |
| H-RPL | Heterosckedastic Random Paramters Logit |
| IID | Independent Identically Distributed |
| MMNL | Mixed-Mixed Multinomial Logit |
| MNL | Multinomial Logit Model |
| NALT | Number of Alternatives |
| NC | No Choice |
| NCT | Number of Choice Tasks |
| NLEV | Number of attributes’ levels |
| OHL | Ordered Heterogeneous Logit |
| ORD | Order |
| PHMNL | Parameterized Heterosckedastic Multinomial Logit |
| RUT | Random Utility Theory |
| S-MNL | Scale Multinomial Logit |
| SQ | Status Quo |
| TC | Task Complexity |
| TR | Transportation |
| WTP | Willingness to pay |

## 1. Introduction

Discrete choice experiments (DCEs) are commonly employed in a wide range of fields, i.e. Marketing, Transportation, Environmental Economics, Health Economics, Tourism, Political Science, originally employed in Psychology (Thurstone 1931). Two are the main building blocks: Random utility Theory (RUT) and Lancaster's Theory of Value (Lancaster 1966). The former was pioneered by the Law of Comparative Judgements (Thurstone 1927), later brought into economics thanks to Marschak (1960). Subsequently, McFadden (McFadden 1974) developed the conditional Logit model, confirming the link between Economics and Econometric theory. This methodology assumes that the respondent follows a compensatory utility maximizing behaviour. Every respondent is assumed to be perfectly rational, not affected by fatigue nor learning, not behaving strategically nor according to different decision processing rules.

Not only has the Psychology literature provided the basis for the discrete choice models based upon RUT, but it has also provided an arena for its critiques. The rationality of human behaviour has been questioned since the 50s (see Simon 1955; Miller 1955). Over time, it has become clear the decision making process was to be regarded as including *'1)cognitive errors that people make when they judge the likelihood of future consequences and 2)simplifying heuristics that people use to cope with the complexity of decision making'* (Loewenstein and Lerner 2003, p.619). Recently, the role of emotions has also received a great deal of attention. The transfer to the DCEs literature has been quite lagged: Araña et al. (2008) have extended the conventional RUT, including the possibility that (stated) emotions play a role as well. In addition, as an alternative to RUT, Chorus et al. (2008) have proposed the Random Regret Minimization approach based on Regret Theory.

Over the last two decades, an increasing number of articles have been focusing on the limitations of the respondents asked to take part in DCEs. In this context, the term Task Complexity (TC) has gained a huge popularity; however, a common definition seems to be lacking. This, in turn, might make practitioners less accountable with regards to the reasons they give for the complexity they set. Indeed, if we are not consistent in defining it, then it appears problematic to be direct and explicit in giving reasons why a certain experimental design has been chosen. We would like the reader not to be left with broad statements only (see Louviere et al. 2011).

For the reasons stated above, a systematic literature review has been conducted using Science Direct, Web of Science, International Bibliography of Social Sciences and Google Scholar. Finally, once defined TC, it appears fundamental to examine how applied DCEs studies are currently dealing with it.

The aim of this paper is threefold: 1) to provide a definition of task complexity in DCEs; 2) reviews the literature concerning those elements defining task complexity; 3) to analyze the recent applied DCEs studies to give evidence of a) which experimental design and dimensions are chosen and b) which motivations are given, if any. The paper's outline mirrors these goals as followed by a section defining the concept of task complexity; Section 3 presents the effects of the task complexity's components defined in Section 2; Section 4 provides a review of the applied work using DCE and finally section 5 concludes.

## 2. Defining task complexity

### 2.1 Individuals and information

The choices we observe are the result of a complex process. First, respondents receive information inputs, next these are processed, and finally the decision is made. Hence, following Mowen (1993), we have three main steps leading to choice:

•*INFORMATION INPUTS*

•*INFORMATION PROCESSING*

•*DECISION MAKING*

Focusing on the information processing level, this in turn consists of exposure, attention and comprehension. All of these are related to involvement and memory. The former can be defined as '*a motivational state influenced by the perceived personal importance and/or interest evoked by a stimulus*' Mowen (1993, p. 73). The more the involvement, the greater the attention and the level of comprehension, namely the extent to which the respondent organizes and interprets the information received. In addition, involvement is generally found to positively influence memory. A worrying issue is that of information overload, which takes place when the respondent receives more information than what he or she can actually process. This may create problems since the respondents may start adopting some heuristics which deviate a lot from the standard assumptions (Malhotra 1984).

## 2.2 What constitutes the task in DCEs

The respondent is presented with a choice experiment which is, to date, usually realized ex ante. The experimental design (ED), so far, has been assessed either by imposing orthogonality between the attributes or seeking the most statistically efficient design. The goal of the former is to ensure the attributes of the design are statistically independent, while the latter aims at producing smaller standard errors for a given sample size. Different measures of the so-called efficient design have been proposed, depending on the assumptions posed on the utility's coefficients and/or the index obtained from the asymptotic variance covariance matrix (AVCM), which in turn depends on the econometric model chosen[3]. Unless one assumes zero or close to zero parameters priors, the orthogonal design is likely to not be the most efficient one whereas an efficient design would most likely have correlations. However, orthogonality in the estimation data is really unlikely to be preserved due to non-response, the removal of response by the analyst, unequal assignment of choice sets, inclusion of covariates in the estimated utility function and finally because non-linear models are estimated[4]. All in all, '*both methods have merits as well as problems'* (Rose and Bliemer 2004, p.2).

Another stream of the experimental design literature is represented by the so called optimal designs, where the Fisher Information Matrix is obtained by taking derivatives against total utility rather than parameters (Street and Burgess 2004a, 2004b, Street et al. 2005). Other constraints on the ED have been also discussed in the literature, such as aiming jointly at utility balance, minimal level overlap, level balance and orthogonality (Huber and Zwerina 1996); minimal attribute overlap in a Bayesian framework (Sandor and Wedel 2001). Crucially, in deciding which approach to follow, the researcher has to consider the quality of the priors available (Ferrini and Scarpa 2007). Recently, developments have sought to go

[3] For a detailed analysis of the state of the art practice see Rose and Bliemer (2004), Rose and Bliemer (2009) and Rose et al. (2009a). Usually, the information criterion relies on the Fisher Information Matrix; see Yu et al. (2012) for alternative information criteria.

[4] This is because the correlations the design seeks to minimize refer to the level of the attributes, but this does not imply there will not be correlations between the differences in the utility function, which is what really matters, as shown by Rose and Bliemer (2004).

beyond an ED where the AVCM is derived with the multinomial Logit model (MNL) as the only data generating process (see Bliemer et al. 2009; Bliemer and Rose 2010). Finally, an approach to handle the model's uncertainty has also been proposed (see Rose et al. 2009a).

A common limitation lies at the heart of each of these EDs. As a matter of fact, concern regarding the limitations of the respondents has been practically non-existent. Two exceptions being blocking the choice situations to reduce the number of choice tasks an individual is asked to take part in or using fractional rather than full factorials designs. One cannot help but notice there has also been an increased interest towards designs that are pivoted around individual-specific reference alternatives (see Rose et al. 2008), which may enhance the realism of the task. However, only a handful of studies (Severin 2001[5]; Sándor and Franses 2009[6]; Danthurebandara et al. 2011[7]) have proposed some ways to take respondent's efficiency into account within the experimental design strategy. If the experimental design has not received much attention from this point of view, most of the reviewed studies have instead focused on what might be the effects arising from different experimental design's dimensions and/or allowing for different processing rules at the estimation stage. In this regard, the main strategy followed has been of relaxing the scale homogeneity assumption, either assuming it exogenously or determining it endogenously, as discussed in the next section.

[5] Specifically, she proposed an overall efficiency index, consisting of '[...] the determinant of the information matrix corrected for the differences in error variance [...]' (Severin, 2001; p.63).

[6] They set the scale factor as a function of the mean attributes level dispersion, the number of trade-offs and price specification.

[7] They propose a D-optimal design where the model assumed is the heteroscedastic logit model, where choice complexity is modelled following Swait and Adamowicz (2001).

There is a crucial link between the experimental design and the econometric models to be estimated after data has been collected. Indeed, the researcher has to determine whether to estimate main effects or non-linear effects too. When he or she aims at maximizing the determinant of the AVCM, the log-likelihood function must be specified, besides setting the parameters values. Notably, a different model implies a different AVCM. Therefore, if a way to incorporate plausible behavioral mechanisms is proven to be empirically valuable at the estimation stage, it should be then transferred, and tested at the experimental design stage.

Irrespective of the experimental design, the researcher must determine:

- (a) the number of choice tasks (NCT);

- (b) the number of alternatives (NALT);

- (c) whether to label or not the alternatives (LAB);

- (d) whether to include a status quo and/or a no choice option (SQ/NC);

- (e) the number of the attributes (NATT);

- (f) the order choice tasks, attributes and alternatives (ORD);

- (g) the number of levels (NLEV);

- (h) the attribute level range (ATT_R);

- (i) the experimental design strategy itself (ED);

- (j) the survey’s administration mode (ADM_M).

## 2.3 Defining Task Complexity

Following Liu and Li (2012), we distinguish between three ways of defining Task Complexity: the structuralist point of view (1), the resource requirement (2) and, finally, the interaction view point (3). According to (1) TC is defined as a function of the elements forming the task and relationships between those elements; (2) looks at the amount of resources a given individual needs in order to complete the task; finally (3) sees TC as the interaction between task and individual's characteristics. As far as DCEs are concerned, following the structuralist point of view we can define objective TC as: *the function of all features and characteristics of the experimental design and the choice experiments' implementation, which have to be set by the researcher. Any of these experimental design's characteristics is likely to influence objective task complexity (TC).*

Formally, the TC of a given choice experiment is modeled as a function of the elements (a) to (j):

$$TC_{CHOICE\ EXP} = f(NCT,\ NALT, LAB, SQNC\ NATT, ORD, NLEV, ATTR, ED, ADMM) + e_i \qquad (1)$$

where the error term $e_i$ reminds there may be some variables influencing $TC_i$ we are not including in (1). For simplicity, we jointly refer to NCT, NALT, NATT, and NLEV as design's dimensions.  Given this level of task complexity, each respondent, characterized by a $i \times k$ matrix of k respondent's characteristics $Z_{ik}$, (such as cognitive ability, involvement, memory) will have a subjective perception of TC in a given choice experiment. Accordingly, focusing on cognitive effort or the time needed by the respondents to complete the survey would fit with the resource requirement view on TC. In regard to the interaction view point, it is given by the interaction between $Z_{ik}$ and $TC_{CHOICE\ EXP}$. In this paper we will focus on the structuralist view of TC and analyze its components in the following sections.

We start from a neutral perspective: more complex choice experiments might be more difficult to be processed by some respondent, but on the other hand they might contribute to a more realistic task and involvement. Task complexity *per se* cannot be deemed to be harmful and necessarily equivalent to difficulty.

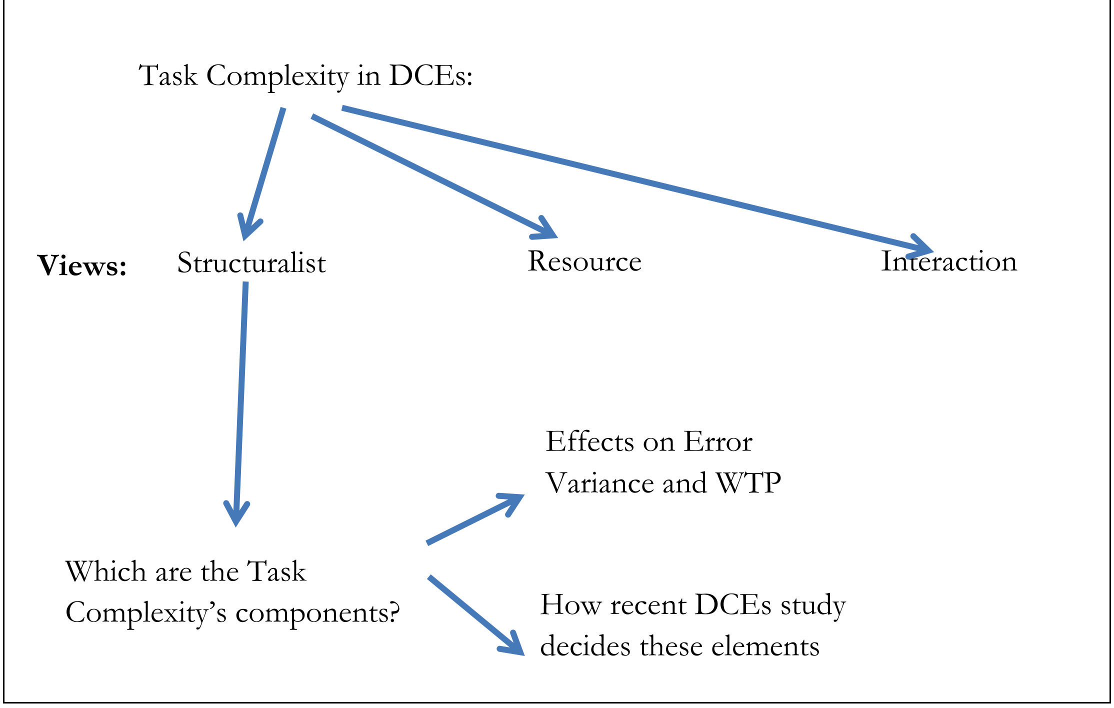
Table 1. Task complexity's definitions

| Views on Task Complexity | Liu and Li (2012) | In DCEs |
|---|---|---|
| **Structuralist** | TC is given by the structure of the task. It can be viewed as a function of the elements forming the task. | TC is given by the function of the experimental design's characteristics and the survey instrument. Together they form the structure of the task. |
| **Resource Requirement** | TC is given as resource requirements needed to complete the task. | TC is given by the cognitive effort, the time needed to complete the task etc. |
| **Interaction** | TC is given by the product of the interaction between the characteristics of the respondents and those of the task. | TC is given by the interaction between task characteristics (entering the function) and respondents' characteristics. |

In other words, complexity cannot be linked only to the quantity of information, but also to its relevance (Hensher 2006). In this sense, the use of perceived difficulty of the task as a proxy for TC is questionable (as in Bonsall and Lythgoe 2009). What is more, the complexity of the task may be linked to different decision strategies (as in Dellaert et al. 2012). However, in this paper we only focus on characterizing TC, whereas linking this with different processing strategies is left for further research.

As shown in Fig.1, after having determined the elements characterizing TC in DCEs following the structuralist view, in the following sections we will be focusing on their effects on error variance and WTP before taking a look at recent DCEs studies.

**Fig.1: Content of the sections 3-4**


Task Complexity in DCEs:
Views:
Structuralist
Resource
Interaction
Effects on Error Variance and WTP
Which are the Task Complexity's components?
How recent DCEs study decides these elements

## 3. Task complexity's components

In the following sub-sections we assess the main findings across different disciplines, firstly considering (a) to (j) in a separate fashion to ease the narrative (with the exception of b-c and g-h treated jointly), before providing a holistic discussion. It is worth noting that only a few authors varied multiple design dimensions and/or presented respondents with different designs in which the main task complexity dimensions have been varied. This seriously undermines the possibility of assessing the impact of design dimensionality on behavioral response (Rose et al. 2009b). Specifically, Dellaert et al. (1999), Swait and Adamowicz (2001), DeShazo and Fermo (2002), Caussade et al. (2005) and Chung et al. (2011) parameterized the scale factor as function of multiple design's dimensions; whereas Hensher (2004, 2006), Cassuade et al. (2005), Rose et al. (2009b), Hess et al. (2012) and Meyerhoff et al. (2013) have presented the respondents with designs with different levels of task complexity (using the so-called Design of Designs). Finally, Rose et al. (2009b) and Hess et al. (2012) obtained data from different countries.

As the scale factor/error variance is a recurrent theme in the following paragraphs, we spend a few lines to give a brief explanation of its role. Heiner (1983) and De Palma et al. (1994) have been the first in linking error variance in random utility models respectively with the gap between respondent's competence and difficulty, and the ability to choose. Hence, a way to assess how the complexity of the choice task affects respondent's behavior is that of focusing on the scale parameter. Indeed, its decrease implies an increase in the variance and vice versa, which has been generally interpreted as less consistent choices. The scale factor can be estimated as setting one data set as the reference (Swait and Louviere 1993) or by setting the scale as a function of some covariates, such as task complexity's dimensions. Different names have been used in the literature to define these models. We use the term heterosckedastic multinomial Logit (HMNL) when scale differences between data sets were estimated. Instead, we employ the term parameterized heterosckedastic multinomial Logit (PHMNL) when the scale factor is a function of same complexity measure(s), hence treating variance heterogeneity in a deterministic way. Besides, the generalized mixed Logit model (G-MNL) has been also proposed (Fiebig et al. 2010), in which random parameters are associated both to the preference coefficients and to scale. However, as warned by Hess and Rose (2012), one cannot claim to completely separate scale and coefficient heterogeneity. Finally, MMNL refers to the multinomial mixed Logit model (sometimes also defined as random parameter Logit model).

### 3.1 Number of choice tasks (NCT)

The first choice experiment ever performed (that we are aware of), consisted of 256 pairwise choice tasks administered to a single respondent (Thurstone 1931). There is no evidence of other studies reaching this number of comparisons. From a theoretical point of view, a single binary choice task is one of the requirements to fulfill incentive compatibility, so that the respondent answers truthfully (Carson and Groves 2007). However, not only does the researcher aim at a receiving a truthful answer, but also to capture the respondents' preferences. On this regard, multiple choice tasks ease the pursuit of this objective, besides allowing for a decrease in data gathering costs and the possibility of estimating interactions (Brazell and Louviere 1996). On this note, we consider for instance Cherchi and Ortúzar (2008) and Rose et al. (2011), which both by means of a simulation analysis, suggest that for a given number of respondents increasing the number of choice tasks produces better t-ratios, as far as MMNL models are concerned. An empirical comparison of the single vs multiple choice format is offered by McNair et al. (2011), who finds differences in WTP arising after the first task. Other authors found the effects on WTP to be small or not significant. This is the case in Chung et al. (2011), Hensher et al. (2001)[8], and Hensher (2004). Finally, Rose et al. (2009b) show how results might be quite culturally and data set specific. In fact, comparing value of travel time savings (VTTS) estimates for Chilean, Australian and Taiwanese respondents, they found the NCT had a significant impact only for the first group of respondents.

The impact on the variance of the error has been receiving a more exacting attention. Bradley and Daly (1994) applied the Logit scaling approach to two case studies, i.e. a series of rank

[8] A preliminary analysis on the same data set, reaching analogous conclusions, was conducted by Stopher and Hensher (2000).

order tasks and a series of choice experiments. Their results show that the scale factors respectively estimated relative to the first rank and the second choice experiment[9] are decreasing over the choices. Therefore, there seems to be evidence, respectively, of a potential rank order effect and a fatigue effect, the latter arising from the 5th response (respondents faced between 10 and 26 choice experiments). Later, Brazell and Louviere (1996) employed a wider range of comparisons, including 12, 24, 48 and 96 choice tasks in one study and 16, 36, 68, 120 in the second. They put forward individuals who seem to learn at first (scale parameter greater than one), whereas some fatigue effect arises at a later stage (scale parameter lower than one)[10]. Similarly, Chung et al. (2011) found a concave relationship between the number of choice sets (1-20) and the scale factor, parameterized as a function of the number of choice sets and alternatives. Interestingly, Cassuade et al. (2005), setting the scale factor as a function of five design dimensions including NCT, confirmed a concave relationship, suggesting that learning effects prevail until the 10th choice task. Nevertheless, this has the least relative impact on variance compared to the other complexity dimensions. Of analogous advice Bech et al. (2011), arguing that although variance is higher for the respondent's segment that had to engage in more choice tasks (17 opposed to 5), the effect appears to be relatively small. The presence of fatigue, and specifically from the 6th task, is also found by Hu (2006).

No statistical evidence of fatigue is present according to Arentze et al. (2003), who estimated the scale of the last 8 choice tasks relative to the first 8. Employing the same comparison as

[9] Since the first was a presented a dominated exercise on purpose, hence discarded as a valid base.

[10] However, differences in variance between the model with 12 and 48 choice sets, case study 1, are not significant. In addition, no scale difference is found across the models with respect to case study 2. Finally, there is no statistical evidence of difference in variance when analysing equivalent sub-design for a given length.

in Arentze et al. (2003), but allowing for heterogeneity in preferences estimating a H-RPL, Carlsson et al. (2012) finds instead evidence of learning effects[11], reversing the evidence of the former study. However, in these two studies the respondents may have realized of taking part in the same sequence of choices twice, so this could have caused order effects. This issue is addressed by Czajkowski et al. (2012), who randomized the order of the choice tasks and estimated HMNL, S-MNL, H-RPL and G-MNL, confirming evidence of learning effects[12]. Finally, no significant effect is found in Meyerhoff et al. (2013) who offered between 6 and 24 choice tasks; but they suggest that higher percentages of drop outs are observed as the NCT grows. Also, no statistical effect is found according to Carlsson and Martinsson (2008), who alternatively present respondents with either 12 or 24 choice tasks, nor by Hole (2004) who asked respondents to take part in 9 choice tasks. Adding up to the not clear effect of the NCT on learning and fatigue, it is important the contribution offered by Hess et al. (2012), who looked for differences in error variance in 5 different case studies in different countries. Although there were differences in scale as one progressed through the choice tasks, a clear and consistent pattern across the studies was not found[13].

Instead using a different approach, namely employing a series of ordered heterogeneous Logit (OHL) models where the dependent variable is the number of attributes ignored, Hensher (2006) suggests that as the number of choice sets increases, from 6 to 15, error variance increases.  Stopher and Hensher (2000) have also wandered outside of the well-

[11] Both Arentze et al. (2003) and Carlsson et al. (2012) set 5 attributes, similar number of attributes levels; the former made use of an orthogonal design, while the latter opted for a D-efficient design.

[12] Each respondent faced 26 choice tasks. According to the HMNL, the scale factor is statistically greater than 1 only after the 10th choice task.

[13] '[...] *overall, there is more evidence of learning (in terms of increasing scale) than fatigue'* (Hess et al. 2012, p. 642).

known path of inspecting error variance. They focused instead on whether increasing the number of tasks leads the respondents to choose the same alternative over and over again, signaling the emergence of fatigue, without finding evidence confirming this hypothesis.

There has been a robust result linked to the number of tasks. Indeed, it has been regularly found that response time decreases as respondents make repeated choices (Haaijer et al. 2000; Rose and Black 2006; Bonsall and Lythgoe 2009; Vista et al. 2009). Coinciding with the evidence that the respondent gets better and better at comparing alternatives, but the learning effect is only one of the possible explanations. For instance, the respondent may just want to get away from the survey.

All in all, there is evidence of a concave relationship between the NCT and the scale factor in those studies where task complexity is not simply seen as a function of NCT and where a great deal of comparisons with respect to the NCT presented (Brazell and Louviere 1996). In regard to the other pieces of evidence, three studies suggest fatigue effects, two put forward learning effects, one suggests no clear pattern, whereas four advance no effect at all. Crucially, when found the effect does not seem to be substantial, neither with respect to the variance of the error nor in terms of mean WTP and elasticities.

### 3.2 Number  of alternatives (NALT)

As far as incentive compatibility is concerned, a binary choice task format would be preferable (i.e.: two alternatives), alongside other conditions to be met (Carson and Groves, 2007). It is arguable whether the binary format might be the most realistic and informative. On this note, some authors have sought to understand whether respondents might prefer more choice opportunities or vice versa. Rolfe and Bennet (2009) asked two groups of respondents to take part, respectively, in a DCE with 2 and 3 alternatives; in each case the status quo option is also available (so effectively we have 3 and 4 alternatives). Their findings suggest that when presented with 4 alternatives, individuals were more able to choose fixed choices and serial non participation decreased drastically. Innovatively, the opportunity of choosing the number of alternatives has been given by Burton and Rigby (2012), specifically offering either 3, 4 or 6 alternatives[14]. Those respondents who opted for 4 or 6 alternatives, in each case 30% of the sample, generally exhibited a lower error variance. For the group with the highest number of alternatives preferred, there was *'a generically articulated desire to have more choices'* (Burton and Rigby 2012, p.794). However, it must be noticed how asking the respondents to choose how many alternatives might increase the complexity of the task, as the respondents are subjected to more and different decision tasks thereby dealing not only with which option to select but how many to consider.

In line with the empirical research on the effects of NCT, many authors have focused on the effects on the error variance of the utility function. Arentze at al. (2003) did not find a statistical significant difference comparing two versus three alternatives (without status quo). Contrary, DeShazo and Fermo (2002) support the evidence of a non-linear relationship, with the scale first increasing and then decreasing. Specifically, error variance is minimized for 3 alternatives (they proposed between 2 to 7 alternatives in one study and 6 to 9 in another study). Similarly, Cassuade et al. (2005) and Mayerhoff et al. (2013) found a quadratic relationship too. As already pointed out, they simultaneously consider other TC dimensions and NALT appears to have the second highest impact. Already in 2001, Swait and

[14] In particular, respondents first received eight choice sets with three alternatives, then one with four and finally one with six options. Eventually, they were asked to choose how many alternatives they would like to have in the next final four choices.

Adamowicz had stressed the relevance of jointly considering multiple TC dimensions. They set the scale factor as a function of an entropy index $H_n$, which is directly affected by NALT, NATT, and N_LEV, putting forward a convex relationship between complexity and error variance. In addition, Chung et al. (2011) found a quadratic relationship between the scale factor and NALT. Notably, they varied the number of alternatives (including a no choice option) from 3 to 12.

Hensher (2004), who did not limit the focus on NALT, in a mixed Logit model where 56 interactions between attributes of alternatives and design characteristics were included, found only 1 statistically significant interaction related to the number of alternatives. Interestingly, as we noted in the previous section for the NCT, also NALT might have a different influence depending on the country considered (Rose et al. 2009b). In fact, in terms of VTTS, it appears that an upward bias is present with respect to the Australian and Taiwanese data set, whereas a downward one when it comes to the Chilean data set.

To sum up, more alternatives are not necessarily associated with a more demanding choice task. For some respondents too few alternatives might undermine their capability of making a choice. Second, with the exception of one study (i.e.: Arentze et al. 2003), it appears to be present a convex relationship between NALT and error variance. This is supported in all those studies setting the scale factor as a function of multiple complexity dimensions, hence less prone to omitted variable bias. Some authors have also indicated the number of optimal NALT (although stressing the contingency of the result), which is interestingly greater than the theoretically incentive compatible binary format.

Notably, researchers have also to decide whether or not to label the alternatives. Usually, the parameters for unlabeled DCE are set as generic, whereas for labelled ones can be either generic or alternative specific. This influences the ED's design dimension, as in the former case less choice situation are required, ceteris paribus. In addition, non-trading behaviour might be associated to labelled alternatives, for instance due to a particular transport mode. However, besides taking into account the risk of fostering non trading behaviour, labeling alternatives may enhance the realism of the task.

### 3.3 Status quo, opt out and no choice option (SQ/NC)

Among the alternatives proposed to the respondents, one should confer keen attention to the inclusion of constant alternatives (CA) in choice tasks, namely the status quo, opt out, neither option. In other words, a crucial decision the researcher has to undertake is whether to design forced or unforced choices. On this note, Pedersen and Gyrd-Hansen (2013) suggest that an unforced DCE needs both an opt out and a status quo option to be undoubtedly such. Unfortunately, there are cases in which both the inclusion and exclusion of a constant alternative are plausible (Pedersen et al. 2011).

If a no choice alternative is included and the respondent chooses it, there is little or no information on the relative attractiveness relative to the other options. In some applied studies, the researcher might incur in a massive number of opt-out choices, hence undermining the possibility of determining the importance of each of the attribute. An example is given by a choice experiment on the realization of nuclear plants in a country where the population is widely against it. To overcome this issue, a dual response choice design may be employed (Dhar and Simonson 2003; Brazell et al. 2006). Accordingly, the respondent is presented with two series of choice tasks, respectively with and without the no-choice option. Similarly, Rose and Hess (2009) propose a dual response format, where in addition to the status quo option is respondent-specific and the alternatives are pivoted with respect to this reference. Moreover, rather than presenting the respondents with two series of choice tasks, they asked them to choose between the remaining alternatives only if the CA had been chosen at first. However, a dual response format may lead to the violation of the independently and identically distributions (IID) assumption, besides differences in WTP estimates. The latter issue has been investigated by Hess and Rose (2009), who do not report systematic differences. Similarly, with respect to the IID violation, Hess and Rose (2009) and Brazell et al. (2006) do not find it to be severe, contrarily to Dhar and Simonson (2003). These authors, however, employed a relatively simple choice experiment compared to Hess and Rose (2009) and Brazell et al. (2006). Differently, Savage and Waldman (2008) propose to first force respondents to choose between the alternatives and later ask them if they would still prefer it to their status quo. All of these approaches aim at possibly tackling a too extreme number of the CA being chosen by first forcing and later setting the respondent free and also choose among nothing or the status quo.

Once it has been established whether to insert a CA in the design, one should also be careful in deciding how to describe it, since this seems to affect respondents' choice too. In this regard, Ruby et al. (1998), in a split sample choice experiment, respectively proposed a *'prefer another site'* option and *'not go fishing'* one, in a context of a saltwater angling survey. They find the way the CA is framed influences results, especially with respect to the '*site characteristics that are most salient to respondents'* (Ruby et al. 1998, p.9). Similarly, Kontoleon and Yabe (2003) alternatively used the *'buying the usual brand'* option and the *'not buying'* one, also concluding that the different format affects the parameters of the attributes that are more important for the respondents. Moreover, the no purchase option appeared to be more restrictive, whereas the usual brand one *'took disproportionately greater share from options that individuals tended to select under the no purchase alternative'* (Kontoleon and Yabe 2003, p.22). Breffle and Rowe (2002) proposed different choice questions formats, i.e. simple choices (2 alternatives), referendum choices (2 alternatives, one is the status quo), and composite choices (multiple changes in attribute levels, no status quo). They argue that, despite the exclusion of the status quo, the simple choice format is associated with the smallest error variance. However, as noticed by the authors themselves, the inclusion of the status quo is not realistic in that instance. All in all, these findings make clear how proposing a status quo rather than an opt out or a no choice option may influence results.

Another issue is that respondents might opt for the CA whilst not stating their true preference. Firstly, because they do not find any alternative appealing enough or because these are too similar, as suggested by Haaijer et al. (2001). Of the same advice Dhar (1997), who suggests the preference for the no choice-option is more likely when an attractive alternative is added to the choice task and less likely when an inferior alternative is added instead. This, in turn, strongly contrasts with Huber and Zwerina (1996) suggestion of aiming at utility balance when designing the choice experiment. On the other hand, Boxall et al. (2009) and Day et al. (2012) suggest that as the difference between the alternatives increases, respondent may find the choice task more difficult and opting out might be viewed as a safe and attractive choice. Similarly, Meyerhoff and Liebe (2009) advance that a greater (perceived) task difficulty may foster the preference for the status quo. Hence, it seems that for some respondents the task gets demanding if too few and similar alternatives are presented, whereas for others the converse may be true. Moreover, varying the number of alternatives might alter the preference towards the no choice or status quo (Adamowicz et al. 2005; Rolfe and Bennet 2009). Another reason why respondents may not answer truthfully is represented by protest against some aspect of the choice tasks (Meyerhoff and Liebe, 2008 and 2009). However, it seems to be particularly challenging to detect protesters in DCE. In this regard, Meyerhoff and Liebe (2008) note that it does not exist a unique way to define protesters when one uses protest/attitudinal scales, so that authors can very likely disagree on whether an individual is deemed to be identified as a protester or not. What is more, respondents might choose the CA because they are uncertain about their preference (Hanley et al. 2006). On this note, Olsen et al. (2011) suggest stated certainty depends positively on the utility difference between alternatives. Finally, Kataria et al. (2012) argue that respondents are more likely to opt out when they find the information to be unbelievable and unrealistic. Realism could be enhanced employing reference based alternatives, although this can lead to high rate of inertia (see Chintakayala et al. 2009).

In a nutshell, once determined whether a CA should be included, a great deal of care has to be placed on the choice of the format and at the modelling stage. In addition, the extent to which alternatives differ seems to have a crucial role in influencing choices, besides task difficulty and protest attitudes. We can safely conclude that status quo and no choice options are not just like any other alternative. As far as TC is concerned, researchers should not simply add up alternatives and CA, but carefully distinguish between the two types.

### 3.4 The number of the attributes (NATT)

As previously noticed with respect to NCT and NALT, it has been also investigated the effect of varying NATT on error variance. Specifically, as the number of attributes increases the error variance seems to increase. This finding is supported by authors who simultaneously considered different complexity dimensions at the estimation stage by means of a PHMNL, as DeShazo and Fermo (2002) and Cassuade et al. (2005). The former varied NATT from 7 to 9, whereas the latter varied them between 4 and 6. In addition, Cassuade et al. (2005) stress that among the complexity dimensions considered, NATT had the largest impact on variance. A substantial impact is also found by Mayerhoff et al. (2013), who varied NATT between 4 and 7. Finally, Swait and Adamowicz (2001) find a concave relationship between the scale factor their entropy index, which is also affected by NATT as noticed in the previous sections.

The conclusions of the authors who only focused on a limited number of complexity dimensions are in line with these results, suggesting this might be quite a robust result, as in the case of Arentze et al. (2003) and Islam et al. (2007). The former considered only NATT and NALT as determinants of task complexity whereas the latter set the scale factor as a function of attributes and attributes' levels. What is more, some attributes can be seen as casually related and not treated independently by the respondents: Blamey et al. (2002) show that the inclusion of a casual attribute reduces the scale factor, while no effect in terms of the taste parameters is found.

Considering instead the effect of NATT on WTP, Rose et al. (2009b) show NATT (varied between 3 and 6) generally influences it downwards, with the exception of two attributes for the Chilean respondents, hence confirming the significance of taking the country and culture factor into account. In addition, Hensher (2006), conditioning the random parameter estimates on stated attribute attendance, finds a lower mean estimate of WTP. Hence, all in all, it appears that both variance and WTP estimates are affected; in addition, respondents may not consider all of the attributes presented.

Specifically, attribute non-attendance-ANA- (Hensher et al. 2005) refers to the possibility that respondents might process only a subset of the attributes in the choice set. As pointed out by Hensher and Rose (2009), mainly two ways have been employed to deal with it: 1) asking respondents supplementary questions to determine whether they are ignoring some attribute (hence using stated ANA) or 2) specifying a model that allows for ANA whilst avoiding the use of supplementary questions (inferring ANA). However, it is worth noticing that ANA should not be of concern when a respondent ignores an attribute because he truly attaches zero, or close to zero, marginal utility to it, hence behaving in a compensatory way (as analogously postulated with respect to respondent deemed to be protesters or not when choosing the status quo-no choice option). In all other circumstances, one should address the issue. Notably, one of these reasons might be task complexity (Hensher 2006; Alemu et al. 2013). However, the possibility that some respondent may only consider a subset of the attributes should not lead practitioners to remove the likely unattended (by some) attributes from the survey. In fact, Hess (2011) suggests that even if for some respondents some attribute might be irrelevant, the risk of overburdening them seems to be small. For recent discussions about ANA, the reader is referred to Alemu et al. (2013), Kehlbacher et al. (2013), Scarpa et al. (2013).

To date, the debate on which method is the most suitable to deal with it (i.e., whether the mixed Logit or latent class framework) and whether to use (and how) or not stated respondents' information on attendance is quite open. The latest evidence suggest that what is being treated as not attended, either by means of stated or inferred approach, might instead only be less important. Hence, practitioners should consider the possibility of asking respondents to state attribute importance, providing a ranking of the attributes at the end of the choice task. In turn, this ranking might be used at the pilot stage to detect seemingly irrelevant attributes.

We conclude this section noticing that among the attributes that need to be chosen to create the scenarios, the cost or monetary one has a crucial role since it allows WTP computation, essential for policy indications. Pedersen et al. (2011) analyze the impact of including/excluding the cost parameter in the context of both forced and unforced choices. In the case of forced situation the scale factor decreases as the cost attribute is included. Moreover, marginal rates of substitution and the ranking of the attributes appear to be affected. The same does not hold in case of unforced choices, where none of these measures seem to be affected. The positive effect in terms of the variance error is also suggested by Carlsson et al. (2007), as well as respondents' preferences[15]. These results show how, as analogously noticed with respect to the alternatives, not every attribute is the same and so not only the number of attributes might affect TC. Further effects are associated also to the way attributes are framed (Howard and Salkeld 2009; Rolfe et al. 2002; Hess et al. 2008a), the context in which decision takes place (Swait et al. 2002; Jaeger and Rose 2008) and whether respondents are familiar with the attributes (Christie and Gibbons 2011).

## 3.5 The number and range of attributes' levels

The effect of varying the number of levels (NLEV) had been examined in the context of conjoint analysis (the reader is referred to Louviere et al. (2010) for a discussion on the differences between choice experiment and conjoint analysis studies).The studies of Currim et al. (1981) and Wittink et al. (1989) suggest that the relative importance of an attribute might increase when this is described by more levels and, according to Wittink et al. (1992), the effect is magnified in the presence of dominated alternatives. Similarly, Green and Srinivasan (1990), suggest that adding more intermediate levels may increase the relative importance of the attribute.

With regard to choice experiments, NLEV might influence the importance attached to the attributes too, but in a different way. According to Hensher (2006) the more the levels per attribute, the less the attributes that may be considered by the respondent. Dellaert et al.

[15] Specifically *'we find one order reversal in each experiment; reject a null hypothesis of equal intensity for several of the preferences (measured in terms of marginal rates of substitutions)'* (Carlsson et al. 2007, p. 162).

(1999) specify the scale parameter as a function of the attribute levels and attribute levels' differences. Both this two have a negative effect on choice consistency and less variability is observed with respect to similarly priced alternatives. Analogously, van der Waerden et al. (2004) suggest that greater variance is attached to four-level attributes as opposed to two-level attributes. As far as the effect on WTP is concerned, Rose et al. (2009) find that the number of level has a significant impact, whose sign is different depending on the attribute and country considered as also noted for NCT and NATT. Similarly, Hensher (2004) found the increase in the number of levels to significantly influence WTP, with different effects depending on the attribute considered. However, regardless of the effects on variance and WTP, practitioners should also keep in mind that more than two levels are needed for a given attribute if non-linear effects are to be estimated (Rose and Bliemer 2009).

Considering the studies which simultaneously model the scale factor as a function of multiple complexity dimensions, Caussade et al. (2005) finds the number of levels contributes to higher variance; however, this effect is three times smaller than the one associated to NATT. DeShazo and Fermo (2002) focus more in depth on changes in the structures of the information provided, analyzing the impact of 1) increasing the number of attributes that differ across alternatives, 2) the mean of the correlation of attribute levels across alternatives and 3) the dispersion of this correlation. Results suggest that an increase in any of these measures increases error variance considerably. Similarly, Sándor and Franses (2009) set the scale factor as a function of mean dispersion (as defined above), the number of trade-offs and price specification. With the exception of the former, the remaining variables affect error variance.  The degree of attribute correlation was also taken into account within the entropy index employed by Swait and Adamowicz (2001), which was found to be convexly related to the error variance.

Not only is important the number of the attributes but also whether wide or narrow ranges are proposed, how much-and how many-levels vary across alternatives and how these are framed. From a statistical point of view a wide range is preferable; nevertheless wide ranges increase the probability of dominated alternatives (Hess and Rose 2009). Empirically, range effects are found to be statistically significant by Cassuade et al. (2005) and Rose et al. (2009). Specifically, the former finds that the levels' range has the third highest impact on variance whereas the latter found significant effects in terms of WTP. Moreover, Hensher (2004) finds that a wide range lowers mean WTP compared to the base and narrower range. In addition, according to Ohler et al. (2000), attribute range differences may impact on the complexity of the functional form, the model fit and between subject response variability. What is more, Luisetti et al. (2011) suggest respondents might interpret the levels as relative rather than absolute values, hence anchoring and referencing may be an issue to be taken into account too.

As previously noted, the price/monetary attribute deserves particular attention. When it comes to the level of the monetary attribute, Hanley et al. (2005) point out that changing the price vector does not seem to affect WTP estimates[16]. On the other hand, Sándor and Franses (2009) suggest that presenting respondents with different (but equivalent) price specification affects choice consistency. Another framing effect can arise depending on whether the levels indicate losses rather than gains. For instance, in a recent study Kragt and Bennet (2012) found that when describing levels in terms of loss, respondents tend to attach more value to it (i.e.: they oppose loss more strongly than they value gains). In order to reduce gains-losses

[16] Nevertheless, the sample presented with higher price levels tends to opt more often for the status quo.

asymmetries, Bateman et al. (2009) stress the importance of presenting the information in the most comprehensible way possible, for example by means of virtual reality visualizations, with the aim of reducing respondents' uncertainty in processing the attributes as this uncertainty may, in turn, lead to an increased preference for avoiding losses.

When it comes to the number of levels differing across alternatives, the study by Mazzotta and Opaluch (1995) ought to be acknowledged. They presented respondents with a minimum of two and a maximum of six attributes' levels differing, concluding that complexity arises, and variance increases, when four or more attributes differ. What is more, Severin (2001) suggests that the number of attribute levels difference is the main responsible for an increase in the error variance. Instead, Maddala et al. (2003) compared a 'minimal overlap' DCE versus a 'increased overlap' DCE, the former characterized by almost no attributes' level overlap, the latter by two overlaps on average. Apart from differences in stated preferences, no differences seem to arise in terms of consistency, dominated responses and perceived difficulty. An analogous study has been later implemented by Johnson et al. (2010), who found statistically different preference's distributions across the two versions of the survey.

We noted in the previous section that respondents might focus on a sub-set of attributes presented. A related issue is that of lexicographic choices. These occur when the respondent chooses on the basis of only one attribute's level, departing from compensatory decision rules (Sælensminde 2006; Campbell et al 2006). Crucially, lexicographic preferences are rational, but fail to satisfy the continuity axiom. In fact, '[...] *for people with lexicographic preferences there does not exist a reservation price at which they are willing to trade a good*.' (Rosenberger et al. 2003, p.64). Hence, the ranges' definitions are crucial: are we reaching respondents' reserve price? Worryingly, lexicographic choices might only apparently be the result of decision strategy to cope with a too difficult task, rather than true underlying preferences (Sælensminde 2006). Practitioners are advised to simultaneously check whether lexicographic choices, attribute non-attendance, non-trading and, more general, inconsistent behaviour are present, as focusing on only one issue seems a rather limited strategy. These checks should be run both at the piloting and model estimation stage and results should be included in the publication.

Summarizing, varying the number of levels seems to affect error variance and WTP. However, it seems rather myopic to only focus on their number. In fact, their range, the way levels are framed, the way they differ between alternatives complicates the picture a lot. Importantly, these effects appear to be robust and substantial, differently from what has emerged when looking at NCT.

### 3.6 The order of choice tasks, attributes and alternatives (ORD)

We have seen in the previous sections how the multiple questions in choice experiments may prompt respondents to anwer in strategic ways, due to its lack of incentive compatibility. Notably, ordering effects might be either linked to the order of the choice task, the attributes and the levels within the choice tasks and, finally, the alternatives.

Providing the same sequence of choice tasks twice, as noticed in the first section, might lead to both learning and fatigue effects. However, we should also reflect upon the effects stemming from presenting respondents with particular sequences of choice tasks and how much information (and how he/she processes it) the respondent is given before starting the experiment. Specifically, Day and Prades (2010) designed their study so that respondents experienced price worsening/improving as well as commodity worsening/improving sequences of choice tasks. Results suggest respondent do not consider each choice task independently and there is evidence of ordering effects with respect to worsening  price and commodity sequences and also with respect to improving commodity sequences. In addition, Day et al. (2012) focused on position-dependent (PO-D) and precedent-dependent (PR-D) ordering effects, respectively related to the position and to *'the nature of the options in the previous task'* (Day et al. 2012, p.74). They analized the impact of informing ex ante respondents about all of the choice tasks' features, as opposed to a task by task revelation of the experiments' characteristics, finding evidence of both PO-D and PR-D ordering effects. The former are significant for the group of respondents who did not received information in advance, while the latter are significant for both the two groups. In particular, PR-D effects are characterized by some anchoring with respect to the first task and to the 'deals' that have been proposed in the previous tasks. Analogously, Scheufele and Bennett (2012) suggest that respondents are affected by the attributes' levels presented in the precedent tasks, finding an increase in cost sensitivity as they progress through the choice tasks.

Inversely, Kjær et al. (2006) examine the effects arising from placing the monetary attribute in different locations in the choice sets (i.e.: at the beginning or at the enf of the attributes listed). Results show the order of the price attribute influences the value respondents attach to it. Particularly, the impact is greater when this attribute is placed at the end of the list. Van der Waerden et al. (2006) focus instead on order effects arising from presenting labelled alternatives in different order. In this case findings suggest significant effects, although not substantial ones.

All in all, order effects are indeed an issue to be faced in choice experiments and, as suggested by Day et al. (2012), we should account for them, further studying the effects of advance disclosure of the choice tasks compared to stepwise task revelation. In addition, it should be further tested whether randomizing choice tasks, attributes and alternatives order across respondents might reduce overall order effects, and at what prize. Finally, in some situations it may be preferable to give respondents the opportunity to order the attributes and alternatives, so to enhance the experiment’s realism (as in Bliemer and Rose 2011).

### 3.7 Experimental design effects

Different strategies have been proposed in order to design choice experiments. However, only a few studies have been conducted in order empirically test whether different ED lead to differences in results. Hess et al. (2008b) compared two orthogonal designs, with and without random blocking, and an efficient design with non random blocking, showing differences in models results and that the random blocking strategy appears to be dominated. Louviere et al. (2008) compared 44 experimental designs in which the number of attributes and the number of attribute levels’ differences were varied. This work stands out since it was aimed at studying the relationship between statistical efficiency and respondents’ choice consistency, beside the great deal of EDs compared. Results show a robust negative relationship between the two measures: the more the attributes and the levels of these, the greater the statistical efficiency and the lower the choice consistency. Previously, Viney et al. (2005) tested the performance of an orthogonal main effects design, a random design and a utility balanced design. Their findings show unexplained variance is greatest for the latter; nevertheless the differences in scale are not significant; overall, the model estimated from the ortoghonal design perform best. Differently, Tudela and Rebolledo (2006) compared the performance of an optimal design which takes into account the variances of the parameter estimates with a classic boundary value one. Although limited by the low number of respondents and incredibly high NCT (16 individuals, 144 choices), hence a non realistic choice experiment, they suggest the former design leads to significantly better t-ratios. Recently, Bliemer and Rose (2011) empirically checked whether D-efficient designs effectively lead to more reliable parameter estimates, given smaller sample sizes. In particular, they compared an orthogonal design and two D-efficient designs, crucially differing in the number of choice situations (108 and 18). Results indeed point towards the use of D-efficient design when the number of choice situation needs to be kept small; nevertheless, their performance relies on the priors the researcher is able to find. In addition, a greater number of choice tasks does not necessarily perform better.

More research is needed to compare ED  which takes into account respondent efficiency (so far focusing on the scale factor as in Severin 2001; Sándor and Franses 2009; Danthurebandara et al. 2011) with the standard ones. Futhermore, more work is needed with respect to the comparison of different experimental design strategies with varying levels of task complexity, as well as in different countries and fields.

### 3.8 Administration mode (ADM_M)

Deciding which survey mode to opt for was of major concern when DCEs were not extensively used yet, whereas Contingent Valuation studies were playing a major role. With respect to this stated preference technique, the NOAA panel (Arrow et al. 1993) advised using the face to face interview method over the mail and telephone, but at a time when the internet was yet to become easily accessible. Face to face interviewing allows for both *'maintaining respondent motivation and allowing use of graphic supplements'* (Arrow et al., 1993; p. 49).

However, the possibility of conducting online surveys, together with the latest development of information technology, gives access to much more powerful graphic representations (see for example Bateman et al. 2009). In addition, it is generally agreed internet surveys are cheaper and returned faster. Besides, they allow the researcher to establish how much and when the given information is displayed. Online surveys can nowadays be shown within the context of face to face surveys, by means of tablet computers, which seem to represent a promising venue for DCEs surveys (see Reiter et al. 2013). However, some segment of the population (for instance, some age group), might find the use of these technologies more difficult or unfamiliar. In addition, these differences may be lessened or enhanced depending on the country considered.  It is worth noting how crucial the choice between a face to face versus a self-completion survey is. As just noted, the interviewer can keep the respondent motivated and observe whether the individual understands the task. But this presence may bias (i.e. due to social desirability and/or satisficing behavior) results, even assuming the interviewer does a great conducting. For instance, Snowball and Willis (2011) show that respondents who self-completed the DCE provided lower-and more realistic- coefficients and WTP estimates. When it comes to task complexity, besides common survey's mode effects[17],

[17] A few studies have focused on sampling and measurements effects. Lindhjem and Navrud (2011) provide a review of survey modes employed in the context of stated preference techniques, concluding there is no clear evidence of differences among the survey modes (nevertheless they recognize most of the study fail to separate sample effects from measurement effects) and that internet surveys might represent the way forward. Windle and Rolfe (2011) compare a paper based format and an internet format using a pre-recruited internet panel. Whereas differences in the sampled groups were found, there were not any in terms of estimated WTP. Bell et al. (2011) focus instead on the sampling effects arising from the use of different recruiting modes. In fact they recruit respondents by phone, in person at malls, online, asking them to complete the survey by means of a computer. They find the latter minimizes self selection issues; besides, the internet recruited sample is associated with significantly fewer inconsistent response (i.e.: choosing dominated alternatives). Analogously, McDonald al. (2010) compare results obtain from online versus mail recruited respondents, concluding that

the administration mode should be chosen and employed so that respondents' understanding and the realism of the tasks are maximized. As regards these concerns, online and/or tablet based survey are essential to carry out an interactive stated choice survey (as in Collins et al. 2012, 2013), which allows respondents to sort, search, show and hide attributes and attributes' levels. In this way each individual is self-selecting the choice set he prefers in order to perform the task. Nevertheless, whereas it is a clearly promising venue for marketing and transportation related studies, it should be tested in the context of environmental and health economics before proclaiming it as a great opportunity for any DCEs study.

The relationship between survey mode and error variance has received some attention in the literature. Savage and Waldman (2008), Olsen (2009) and Hatton McDonald et al. (2010) investigated whether the survey mode affects choice consistency, finding that respondents answered more consistently in the mail survey rather than in the internet one. Previously, Brydon (1997) compared paper and pencil versus computer interactive surveys, varying the number of alternatives and attributes presented, concluding that paper-pencil format is preferable when NATT and NALT are relatively high (i.e. six alternatives and twelve alternatives). However, studies are needed where task complexity, as in this paper defined, is analyzed in relation to different survey modes. It would be especially interesting to investigate whether the combination of online and the latest IT developments available make the respondent more involved and/or make choices more consistent.

those mail recruited and surveyd have the greatest error variance, whereas the lowest is for those who were internet recruited.

## 3.9 A summary of the findings

Although the studies reviewed so far present a low degree of comparability it is fair to draw the following conclusions:

- First, there is evidence that each choice situation is unlikely to be independent from the previous one for reasons including learning, fatigue and ordering effects. In addition, results indicate an increased number of choice tasks seem to affect error variance and estimated willingness to pay. Nevertheless, these effects are consistently reported to be small/not substantial.

- Second, as regard the number of alternatives, evidence points towards a concave relationship with the scale factor, primarily in those studies that simultaneously considered multiple task complexity's dimensions. Crucially, practitioners have to carefully consider whether to include a status quo/no choice option and how to frame it. Some studies have reported that a more difficult choice task can lead to an increased preference for the status quo/no choice.

- Third, as far as the number of the attributes is concerned, results suggest that error variance is substantially and positively affected. This conclusion is rather robust across the studies reviewed. Furthermore, there is some evidence of effects in terms of the willingness to pay.

- Fourth, as the number of levels increase, error variance and willingness to pay seem again to be affected. Notably, non-linear effects on scale are consistently reported by authors who parameterized it as a function of multiple measures linked to the number of levels (such as mean and standard deviation of the correlation between attributes' levels). Moreover, varying levels' range may impact variance and WTP as well.

- More research is needed to assess the effect of the experimental design strategy. Particularly, more real data studies are needed rather than simulations, in order to test how respondents perform in an experimental design set up considering respondent's efficiency. An interesting result is that a greater number of choice tasks does not necessarily performs better, and a higher number of attribute and attributes' levels is associated with less choice consistency.

- Finally, online/interactive surveys seem to represent an effective way to foster realism and respondents' involvement. Nevertheless, in some study reports respondents seem to answer more consistently in mail surveys. Research is needed to assess how task complexity is perceived in relation to different survey modes.

## 4. Task complexity's awareness in the applied literature

We selected 30 recent (2011-2013) applied studies, ten each from Transportation (TR), Environmental Economics (ENV) and Health Economics (HE) literature. Specifically, these studies have been selected from Transportation Research, Transportation, Ecological Economics, Land Economics, Energy Policy, Journal of Health Economics, Health Economics and Health Policy. Table 2 shows the TC dimensions chosen in these studies.

We begin with providing some summary statistics. First, we notice some variability in terms of NCT. The minimum number observed is 3, the maximum is 32, whereas the average number of choice tasks presented is 9. The NCT is higher in the HE studies, whereas is quite similar if compared between TR and ENV. Second, most of the studies present between two and three alternatives; an exception being Ziegler (2012) with 7. Notably, within the ENV applied studies, it is frequently prevalent the availability of the status quo or an opt out alternative. Almost all of the studies present unlabeled alternatives. Third, the average number of attributes chosen is around 6 for the TR and HE applied studies, whereas it is less than 5 for the ENV papers. More dispersed is the picture within the TR literature, where three studies use as much as 9 attributes, while one as few as 3. Fourth, the attributes' levels range between 2 and 8; only in one case one attribute can have up to 16 levels (in Viney et al. 2013). Fifth, when it comes to the experimental design, the main criteria still employed are orthogonality and efficiency. Unfortunately, some studies do not provide a detailed explanation regarding their ED. None of the studies considered applies the EDs proposed by Severin (2001), Sándor and Franses (2009) or Danthurebandara et al. (2011) in order to deal with respondent efficiency.  Finally, considering the survey instrument, we notice that face to face interviews seem to be rarely used within TR and HE, whereas the online surveys are gaining more popularity, especially within the TR studies.

There are some substantial differences, in terms of task complexity dimensions, set across the studies considered. For instance, Popkin et al. (2012) choose 3 alternatives, 3 attributes and 4 choice tasks, whereas Fiebig et al. (2011) set them respectively to 3, 8 (with 2 of these attributes having 8 levels) and 32. Are these differences influenced by the authors' beliefs about respondents' cognitive burden or are they set following a specific protocol? Unfortunately, only in a few of these studies the authors explicitly refer to this issue, failing to explain why they chose those particular design's dimensions. Most of them focus on the number of the choice tasks. For instance, in Gracia et al. (2012, p.788) we find that 'To avoid fatigue effects associated with multiple scenario valuation tasks, the 32 choice sets were randomly split into 8 blocks of 4 choices.' Similarly, Hurley and Mentzakis (2013, p.674) suggest 'To reduce the burden on subjects the 24 choice scenarios were divided into two blocks of 12 scenarios […]'. Mentzakis et al. (2011, p.934) put forward that 'The discussion groups and pilot surveys indicated that respondents found it difficult to engage in 48 choices.

The design was therefore blocked into six sets of eight choice sets'. Nguyen et al. (2013, p.120) state that 'In order to reduce the cognitive burden on respondents, each respondent was randomly chosen to face a block of 6 choice tasks'. Finally in Kolstad (2011, p.200) it is reported that 'The 32 choices were randomly divided into two blocks in order not to exhaust the respondents'.

Only a few studies have explicitly expressed concern over the number of attributes. For example, Duke et al. (2012, p.98) state that '[…] the number of attributes is at the mid-to-lower end of most choice experiments, leading to relatively simple choice task'. Similarly, in Kolstad (2011, p.200) '[…] seven attributes are included. Thus, the job alternatives include a relatively complete description of the job, while at the same time they avoid being too complex for rational and well-informed choice making'. Some attention is paid to order effects issues, namely in Nguyen et al. (2013), Kolstad (2011), Hurley and Mentzakis (2013), Franken and Koolman (2013) and Popkin et al. (2012). However, this is simply addressed by varying the order of choice tasks presented to the respondents.  The remaining TC elements are mainly stated with no concerns about the respondents.

In some of these studies, focus groups and in-depth interviews have been used to help determine NCT and NATT. However, it is not clear what procedure has been used, hence not giving more insights than authors simply stating what was the NCT and NATT chosen. For instance, how do we pass from 48 to 8 choice sets (in Mentzakis et al. 2011)? Furthermore, it appears to be uncommon to use focus groups and/or interviews to build the scenarios and test the feasibility of *higher* levels of objective TC, instead of focusing almost exclusively on the cognitive limits of the respondent. Moreover, the cognitive burden is mainly associated to the number of choice tasks. But we have just concluded in the previous section how this might be over-feared.

In conclusion, the issue of task complexity and cognitive burden is acknowledged by the authors in the building of the choice experiments. However, we confirm 1) there is not a clear definition of these concepts, 2) too little information is given regarding the NCT and NATT, 3) almost none of the studies gives reasons why the other TC dimensions have been chosen.

Table 2.Applied DCEs studies

| Journal | Reference | Alternatives | Attributes | Levels | NCT | Sample size | Obs. | Model | Labelled/Unlabelled | Type | Survey mode | CE construction |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T Res A | Beck et al. (2013) | 4 (1 status quo) | 9 | 3,6 | 5 | 650 | 3172 | LC | Labelled | D-efficient | Online+interviewer | Literature reviews, secundary data analysis |
| T Res A | Jones et al. (2013) | 3 | 9 | 1(2) 2(3) 2(6) 2(4) 2(5) | 6 | 400 | 2236 | RPL | Labelled | Orthogonality | Face to face | Literature review, interviews |
| Transp | Rose and Hensher (2013) | 3(1 status quo) | 7 | 4(4) 3(3) | 12 | 189 and 295 | 2268 and 3540 | ECL | Unlabelled | D-optimality | online | Previous study experience |
| Transp | Sikka and Hanley (2013) | 2 or 3(1 status quo) | 4 | 2(6) 1(4) 1(5) | 12 | 273 | 2088 | RPL | Unlabelled | Fractional Factorial, pivoted alternatives | online | Literature review |
| T Res D | Caulfield et al. (2012) | 4,6 | 5 | 3(3) 1(4) 1(2) | n.a. | 1941 | 11692 | MNL | Unlabelled | Fractional Factorial | CAPI | Specific |
| T Res A | Devarasetty et al. (2012) | 4 | 3 | 3,[a]RGWI | 3 | n.a. | 3325 | RPL | Unlabelled | Db eff; random, adaptive random | Online | Literature review |
| T Res A | Wardman and Ibánez (2012) | 2 | 6 | 5,8 | 9 | 956, 1040 | 8426, 9359 | Het-Logit | Unlabelled | Fractional Factorial | 1 mail, 1 online | Literature review |
| T Res A | Ziegler (2012) | 7 | 5 | 4(5) 1(4) | 6 | 598 | 3588 | M Probit | Quasi-labelled | n.a. | CAPI | Pretest |
| T Res A | Correia and Viegas (2011) | 2 | 8,9 | [a]RGWI | 4 | 996 | 3984 | Binary Logit | Labelled | Fractional Factorial | Online | Past experience, literature review |
| T Res D | dell'Olio et al. (2011) | 2 | 4 | 4(3) | 9 | 110 | 990 | RPL | Unlabelled | Balanced, not further specified | interviewed | Literature review, FG, pilot |

Table 2. Continued

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ECEC | Abildtrup et al. (2013) | 3(1 status quo) | 5 | 5(3) | 6 | 1054 | 6324 | RPL | Unlabelled | $D_b$ efficient | Online | Literature review, FG, interviews, pilot |
| ECEC | Nguyen et al. (2013) | 2(1 status quo) | 4 | 3(3) 1(2) | 6 | 169 | 1014 | MNL, RPL, LC | Unlabelled | Orthogonality | Face to face | Pilots, internet survey, interviews, literature review |
| ECEC | Gelo and Koch (2012) | 3(1 status quo) | 4 | 2(2) 1(4) 1(5) | 4 | 600 | 2400 | MNL, RPL, LC | Unlabelled | Orthogonality | Face to face | FG, Meta analyis, pilot |
| ECEC | Duke et al. (2012) | 3(1 status quo) | 5 | 1(4) 3 (2) 1(5) | 5 | 664 | 3280 | RPL | Unlabelled | D-optimality | Mail | FG, Pretest |
| Land Econ | Anderson et al. (2013) | 3(1 opt out) | 8 grouped into 4 | 1,3 | 3 | 1309 | 4752 | RPL | Unlabelled | D-efficient | Mail | FG, interviews |
| EP | Kaenzig et al. 2013 | 3 | 7 | 5(4) 2(5) | 12 | 414 | 4968 | HBM | Unlabelled | Random | CAPI | Literature review and expert interviews |
| EP | Kosenius and Ollikainen (2013) | 4 5(1 opt out) | 4 | 3(8) 1(4) | 8 | 947 | 7566 | NL | Unlabelled | Orthogonality | Online | Literature review, interviews, pilot |
| EP | Gracia et al. (2012) | 3(1 status quo) | 5 | 4(5) 1(2) | 4 | 400 | 1600 | RPL | Unlabelled | Orthogonality | Face to face | Literature review, interviews, FG pilot |
| EP | Popkin et al. (2012) | 3(1 status quo) | 3 | 2(2) 1(7) | 4 | 515 | 1425 | RPL | Unlabelled | Random | Online | Literature Review |
| Land Econ | Qin et al. (2011) | 3(1 opt out) | 5 | 1(5) 2(2) 2(3) | 7 | 210 | 1470 | RPL | Unlabelled | Orthogonality | Face to face | FG, interviews, pilot |

Table 2. Continued

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| J Health Ec | Hurley and Mentzakis (2013) | 4(1 opt out)/2(1 opt out) | 6,4 | 5(2) 1(6)/3(2) 1(6) | 12,16 | 268 | 3216/4288 | Nested Logit | Unlabelled | Orthogonality | Online | Literature Review |
| Health Pol | Franken and Koolman (2013) | 2 | 5 | 4 | 16 | 63 | 1008 | MNL | Unlabelled | Orthogonality | Pencil paper | Pilot, involvment pre-survey |
| Health Econ | Viney et al. (2013) | 3 | 6 | 5(3) 1(16) | 15 | 1031 | 15465 | QALY | Unlabelled | Specific | online | Literature review, Pilot |
| Health Pol | Goodall et al. (2012) | 2 | 6 | 4(2) 2(4) | 16 | 78 | 1247 | MNL, RPL | Unlabelled | Full factorial randomly splitted | not stated | Literature review |
| Health Pol | Pederesen et al. (2012) | 3(1 status quo) | 4 | 2(5) 1(2) 1(4) | 4 | 1229 | 4916 | MNL | Labelled | $D_b$ efficient | Mail | Interviews, pilot |
| Health Econ | Sivey et al. (2012) | 2 | 7 | 6(3) 1(4) | 9 | 536 | 4808 | GMNL | Unlabelled | D error | paper or online | Interviews, pilot, literature review |
| Health Pol | Watson et al. (2012) | 3(1 status quo) | 5 | 4(4) 1(2) | 11 | 450 | 4950 | MNL | Unlabelled | Orthogonality | Mail | Literature review, interviews, discussions |
| Health Econ | Fiebig et al. (2011) | 3 | 8 | 1(4) 2(8) 2(4) 3(3) | 32 | 171 | 5472 | MNL RPL GMNL | Unlabelled | Modified LMA | online | Literature review, interviews, discussions. FG |
| Health Econ | Kolstad (2011) | 2 | 7 | 3(4) 4(2) | 16 | 296 | 9342 | MNL | Unlabelled | Orthogonality | Face to face | Literature review, interviews, discussions |
| Health Econ | Mentzakis et al. (2011) | 3( 1 opt out) | 5 | 5(4) | 8 | 209 | 1672 | LC | Unlabelled | D optimal | Mail | Literature review, Pilot |

[a] RGWI: Randomly Generated between Intervals. Journals' abbreviations: ECEC=Ecological Economics. EP=Energy Policy. Health Pol=Health Policy. J Health Ec= Journal of Health Economics. Land Econ=Land Economics. T RES= Transportation Research. Transp=Transportation. Models' abbreviations: M Probit=Multivariate Probit. HBM= Hierarchical Bayes Model. Other: FG=Focus Group.

## 5. Conclusions

We have sought to define task complexity within the choice experiments method, following the structuralist angle. We have also thoroughly reviewed published articles that studied the effect of TC dimensions on error variance and WTP. Finally, we have considered 30 recently published articles wondering if and how this issue has been addressed. Although this theme has gained importance, it appears there is no protocol established so to justify the task complexity dimensions' chosen. Notably, the publishing process might also influence the amount of information given.

Task complexity is a small component of a bigger issue. Respondents might develop numerous strategies during the experiment, such as lexicographic, attribute non-attendance, non-trading and mixings of these. Hence, the least practitioners can do is to check for heteroskedasticity and correlations in the data and test more complex models, already available, which can take into account different heuristics (see for example Araña et al. 2008, Leong and Hensher 2012). In addition, it is not just about the number of choices, attributes, attributes' levels and alternatives. Not all attributes are the same (i.e.: the cost attribute), not all alternatives are the same (i.e.: status quo/no choice). If this is not enough, influences and effects seem to be rather country and culturally specific.

Hence, unfortunately, it does not seem possible to provide guidelines concerning the number of each one of these dimensions. Frankly, any rules of thumb in this context seem rather naive and studies in which some numbers are provided should be looked at cautiously and critically. Surely, a great deal of task complexity awareness is needed, and this is what we call for: to step into the respondent's shoes. What all the studies reviewed here tell us is that a higher level of objective complexity is not necessarily abominable for the respondents. Asking whether they are able to understand the task and/or what happens if it is too difficult for them is not enough. Realism, as repeatedly stressed in the works by Louviere and Hensher, is simply essential and can be fostered by tailoring the choice task to the context, the country, the respondent herself. On this note, fixed experimental design's dimensions, i.e. the same for each respondent, might be dominated by an ED with variable dimensions, which can be also dynamically adjusted during the choice experiment, relying on online-interactive surveys. Research is needed to assess this. On this note, individual efficient experimental designs might represent a great opportunity (see Rose and Hensher 2013).

In light of these considerations, a careful qualitative analysis to inform the survey's construction appears to be a necessary prerequisite, *'a must',* as already stated by Hess and Rose (2009). But at the same time, some set of rules should be established and empirically tested: broad statements regarding the setting of the ED should be taken over by detailed reasons. This appears to be fundamental to get to know more about how to practically deal with the issue and provide examples of 'good practice' for the practitioners to come, which cannot overly rely on their 'own judgment'.

The research has much more to investigate. We need cross-field studies to determine the extent to which the state of the art tools, and whether further extensions, in Transportation and Marketing are transferable to Health and Environmental Economics. Another notable gap in the literature is assessing how different ED strategies perform as ED dimensions are varied. In addition, research is needed to test whether extending the averaging approach framework proposed by Rose et al. (2009a) to include models that take into account heterosckedasticity and/or different processing strategies can likely lead to substantial gains and if so, in which circumstances. Finally, for the proponents of the value transfer technique, one should critically interrogate if and how to consider task complexity when transferring values.

## References

Abildtrup J, Garcia S, Olsen S B, Stenger A (2013) Spatial preference heterogeneity in forest recreation. Ecological Economics 92:67-77

Adamowicz W, Dupont D, Krupnick A (2005) Willingness to pay to reduce community health risks from municipal drinking water, a stated preference study. WP, Department of Rural Economy, University of Alberta

Adamowicz W, Bunch D, Cameron T A, Dellaert B G C, Hanneman M, Keane M, Louviere J, Meyer R, Steenburgh T, Swait J (2008) Behavioral frontiers in choice modeling. Marketing Letters 19:215-228

Alemu M H, Mørbak M R, Olsen S B, Jensen C L (2013) Attending to the reasons for attribute non-attendance in choice experiments. Environmental and Resource Economics 54:333-359

Anderson L E, Todd Lee S, Levin P S (2013) Costs of delaying conservation: Regulations and recreational values of exploited and co-occurring species. Land Economics 89: 371-385

Araña JE, León CJ, Hanemann MW (2008) Emotions and decision rules in discrete choice experiments for valuing health care programmes for the elderly. J Health Econ 27:753–69

Arentze T, Borgers A, Timmermans H, Del Mistro R (2003) Transport stated choice responses: effects of task complexity, presentation format and literacy. Transportation Research Part E 39:229-244

Arrow K, Solow R, Portney P R, Leamer E E, Radner R, Schuman H (1993) Report of the NOAA Panel on Contingent Valuation

Bateman I J, Day B H, Jones A P, Jude S (2009) Reducing gain-loss asymmetry: A virtual reality choice experiment valuing land use change. Journal of Environmental Economics 58(1): 106-118

Bech M, Kjaer T, Lauridsen J (2011) Does the number of choice sets matter? Results from a web survey applying a discrete choice experiment. Health Economics 20:273-286

Beck M J, Rose J M, Hensher DA (2013) Environmental attitudes and emissions charging: An example of policy implications for vehicle choice. Transportation Research A 50:171-182

Bell J, Huber J, Viscusi W K (2011) Survey mode effects on valuation of environmental goods. International Journal of Environmental Research and Public Health 8:1222-1243

Blamey R K, Bennet J W, Louviere J J, Morrison M D, Rolfe J C (2002) Attribute casuality in enviromental choice modelling. Environmental and Resource Economics 23:167-186

Bliemer M C J, Rose J M, Hensher DA (2009) Efficient stated choice experiments for estimating nested Logit models. Transportation Research Part B 43(1):19-35

Bliemer M C J, Rose J M (2010) Construction of experimental designs for mixed Logit models allowing for correlation across choice observations. Transportation Research B 44(6):720-734

Bliemer M C J, Rose J M (2011) Experimental design influences on stated choice outputs: An empirical study in air travel choice. Transportation Research Part A 45:63-79

Bonsall P, Lythgoe B (2009) Factors affecting the amount of effort expended in responding to questions in behavioural choice experiments. Journal of Choice Modelling 2(2):216-236

Boxall P, Adamowicz W L, Moon A (2009) Complexity in choice experiments: choice of the status quo alternative and implications for welfare measurement. The Australian Journal of Agricultural and Resource Economics 53:503-519

Bradley M, Daly A (1994) Use of the Logit scaling approach to test for rank-order and fatigue effects in stated preference data. Transportation 21:167-184

Brazell J D, Diener C G, Karniouchina E, Moore W L, SeverinV, Uldry P F (2006) The no choice option and dual response choice design. Marketing Letters 17: 255-268

Brazell J D, Louviere J J (1996) Length effects in conjoint choice experiments and surveys: An explanation based on cumulative cognitive burden. Working Paper- Dept. of Marketing-University of Sydney

Breffle W S, Rowe R D (2002) Comparing question formats for evaluating natural resources trade-offs. Land Economics 78 (2): 298-314

Brydon K (1997) Demand effects in experimental choice: Task complexity and the method of administration. The University of Sydney Marketing Honours thesis

Burton M, Rigby D (2012) The self selection on complexity in choice experiments. American Journal of Agricultural Economics 94 (3):786-800

Cairns J, Van Der Pol M (2004) Repeated follow-up as a method for reducing non-trading behaviour in discrete choice experiments. Social Science and Medicine 58:2211-2218

Carlsson F, Frykblom P, Lagerkvist C J (2007) Preferences with and without prices-does the price attribute affect behaviour in stated preference survey? Environmental and Resource Economics 38:155-164

Carlsson F, Martinsson P (2008) How much is too much? An investigation of the effect of the number of choice sets, context dependence and the choice of bid vectors in choice experiments. Environmental and Resource Economics 40:165-176

Carlsson F, Mørbak M R, Olsen S B (2012) The first time is the hardest: A test of ordering effects in choice experiments. Journal of Choice Modelling 5(2):19-37

Carson R T, Groves T (2007) Incentive and informational properties of preference questions. Environmental and Resource Economics 37:181-210

Carson R, Louviere J J, Anderson D A, Arabie P, Bunch D S, Hensher D A, Johnson R M, Kuhfeld, W, Steinberg D, Swait J, Timmermans H, Wiley J B (1994) Experimental analysis of choice. Marketing Letters:351-368

Cassuade S, Ortùzar J de D, Rizzi L I, Hensher D A (2005) Assessing the influence of design dimension on stated choice experiment estimates. Transportation Research Part B 39:621-640

Caulfield B, Brick E, McCarthy O T (2012) Determining bicycle infrastructure preferences-A case study of Dublin. Transportation Research D 17:413-417

Cherchi E, Ortúzar, J de D (2008) Empirical identification in the mixed Logit model: Analysing the effect of data richness. Network and Spatial Economics 8:109-124

Chintakayala P K, Hess S, Rose J M (2009) Using second preference choices in pivot surveys as a means of dealing with inertia. In: European Transport Conference, Leeuwenhorst, October 5-7

Chorus CG, Arentze T A, Timmermans HJP (2008) A Random Regret-Minimization model of travel choice. Transp Res Part B Methodol 42:1–18

Christie M, Gibbons J (2011) The effect of individual ‘ability to choose’ (scale heterogeneity) on the valuation of environmental goods. Ecological Economics 70:2250-2257

Chung C, Boyer T, Han S (2011) How many choice sets and alternatives are optimal? Consistency in choice experiments. Agribusiness 27 (1):114-125

Collins A T, Hess S, Rose JM (2013) Choice modelling search and sort data from an interactive choice experiment. Transportation Research Part E 56:36-45

Collins A T, Rose JM, Hess S (2012) Interactive stated choice surveys: A study of air travel behavior. Transportation 39(1):55-79

Correia G, Viegas J M (2011) Carpooling and carpool clubs: clarifying concepts and assessing value enhancement possibilities through a Stated Preference web survey in Lisbon, Portugal. Transportation Research A 45:81-90

Currim I S, Weinberg C B, Wittink D R (1981) Design of subscription programs for a performing art series. Journal of Consumer Research 8:67-75

Czajkowski M, Giergiczny M, Greene W (2012) Learning and fatigue effects revisited. The impact of accounting for unobservable preference and scale heterogeneity on perceived ordering effects in multiple choice task discrete choice experiments. Faculty of Economic Sciences, University of Warsaw-WP N 8/2012

Danthurebandara VM, Yu J, Vanderbroek M (2011) Effect of choice complexity on design efficiency in conjoint choice experiments. Journal of Statistical Planning and Inference 141: 2276-2286

Day B, Bateman I, Carson R T, Dupont D, Louviere J, Morimoto S, Scarpa R, Wang P (2012) Ordering effects and choice set awareness in repeated response stated preference studies. Journal of Environmental Economics and Management 63:73-91

Day B, Prades J L P (2010) Ordering anomalies in choice experiments. Journal of Environmental Economics and Management 59:271-285

De Bekker-Grob E W, Hol L Donkers B, Van Dam L, Habbema J D, Van Leerdam M E, Kuipers E J, Essink-Bot M L, Steyerberg E W (2010) Labeled versus unlabeled discrete choice experiments in health economics: An application to colorectal cancer screening. Value in Health 13 (2):315-323

Dellaert B G C, Brazell J D, Louviere J J (1999) The effect of attribute variation on consumer choice consistency. Marketing Letters 10 (2):139-147

Dellaert B G C, Donkers B, Van Soest A (2012) Complexity effects in choice experiment-based models. Journal of Marketing Research 49:424-434

Dell'Olio L, Ibeas A, Cecín P, dell’Olio F (2011) Willingness to pay for improving service quality in a multimodal area. Transportation Research part C 19:1060-1070

De Palma A, Myers G, Papeorgiou Y (1994) Rational choice under imperfect ability to choose. American Economic Review 84:419-440

DeShazo J R, Fermo G (2002) Designing choice sets for stated preference methods: The effects of complexity on choice consistency. Journal of Environmental Economics and Management 44:123-143

Devarasetty P C, Burris M, Douglass Shaw W (2012) The value of travel time and reliability-evidence from a stated preference survey and actual usage. Transportation Research Part A 46:1227-1240

Dhar R (1997) Consumer preference for a No-choice option. The Journal of Consumer Research 24 (2):215-231

Dhar R, Simonson I (2003) The effect of forced choice on choice. Journal of Marketing Research 40 (2):146-160

Duke J M, Borchers A M, Johnston R J, Absetz S (2012) Sustainable agricultural management contracts: Using choice experiments to estimate the benefits of land preservation and conservation practices. Ecological Economics 74:95-113

Ferrini S, Scarpa R (2007) Design with a priori information for nonmarket valuation with choice experiments: A Monte Carlo study. Journal of Environmental Economics and Management 53:342-363

Fiebig D G, Knox S, Viney R, Haas M, Street D J (2011) Preferences for new and existing contraceptive products. Health Economics 20:32-52

Franken M, Koolman X (2013) Health system goals: A discrete choice experiment to obtain societal valuations. Health Policy 112:28-34

Gelo D, Koch S F (2012) Does one size fit all? Heterogeneity in the valuation of community forestry programs. Ecological Economics 74, 85-94

Goodall S, King M, Ewing J, Smith N, Kenny P (2012) Preferences for support services among adolescents and young adults with cancer or a blood disorder: A discrete choice experiment. Health Policy 107:304-311

Gracia A, Barreiro-Hurlé J, Pérez y Pérez L (2012) Can renewable energy be financed with higher electricity prices? Evidence from a Spanish region. Energy Policy 50:784-794

Haaijer R, Kamakura W, Wedel M (2001) The ‘no-choice’ alternative in conjoint choice experiments. International Journal of Market Research 43:93-106

Haaijer R, Wagner K, Wedel M (2000) Response latencies in the analysis of conjoint choice experiments. Journal of Marketing Research 37:376-382

Hanley N, Adamowicz W, Wright RE (2005) Price vector effects in choice experiments: an empirical test. Resource and Energy Economics 27:227-234

Hanley N, Wright R E, Alvarez-Farizo B (2006) Estimating the economic value of improvements in river ecology using choice experiments: an application to the water framework directive. Journal of Environmental Management 78:183-193

Hatton McDonald D, Morrison M, Rose J M, Boyle K (2010) Untangling differences in values from internet and mail stated preferences studies. IV World Congress of Environmental and Resource Economists, Montreal

Heiner R (1983) The origin of predictive behaviour. American Economic Review 73:560-595

Hensher D A (2004) Accounting for stated choice design dimensionality in willingness to pay for travel time savings. Journal of Transport Economics and Policy 38:425-446

Hensher D A (2006) How do respondent process stated choice experiments? Attribute consideration under varying information load. Journal of Applied Econometrics 21:861-878

Hensher D A, Rose J M, Greene W (2005) The implications on willingness to pay of respondents ignoring specific attributes. Transportation 32 (3):203-222

Hensher D A, Rose J M (2010) Simplifying choice through attribute preservation non-attendance: Implications for willingness to pay. Transportation Research Part E 45:583-590

Hensher DA, Stopher P R, Louviere J J (2001) An exploratory analysis of the effect of number of choice sets in designed choice experiments: an airline choice application. Journal of Air Transport Management 7:373-379

Hess S (2011) Impact of unimportant attributes in stated choice surveys. 91th Annual Meeting of the Transportation Research Board TRB, Washington DC, US

Hess S, Hensher D A, Daly A (2012) Not bored yet-Revisiting respondent fatigue in stated choice experiments. Transportation Research Part A 46:626-644

Hess S, Rose J M (2009) Some lessons in stated choice survey design. In: European Transport Conference, Leeuwenhorst, October 5-7

Hess S, Rose J M, Hensher D A (2008a) Asymmetric preference formation in willingness to pay estimates in discrete choice models. Transportation Research Part E 44:847-863

Hess S, Rose J M, Polak J (2010) Non-trading, lexicographic and inconsistent behaviour in stated choice data. Transportation Research Part D 15:405-417

Hess S, Smith C, Falzarano S, Stubits J (2008b) Managed-Lanes stated preference survey in Atlanta, Georgia. Measuring effects of different experimental designs and survey administration methods. Transportation Research Record: Journal of the Transportation Research Board, No. 2049:144-152

Hole A (2004) Forecasting the demand for an employee Park and Ride service using commuters' stated choices. Transport policy 11:355-362

Horowitz J K, McConnell K E (2002) A review of WTA/WTP studies. Journal of Environmental Economics and Management 44:426-447

Howard K, Salkeld G (2009) Does attribute framing in discrete choice experiments influence willingness to pay? Results from a discrete choice experiment in screening for colorectal cancer.  Value in Health 12 (2):354-363

Hu W (2006) Effects of endogenous task complexity and the endowed bundle on stated choice. American Agricultural Economic Association Annual Meeting, Long Beach, California-July 23-26

Huber J, Zwerina K (1996) The importance of utility balance in efficient choice designs. Journal of Marketing Research 33:307-317

Hurley J, Mentzakis E (2013) Health-related externalities: Evidence from a choice experiment. Journal of Health Economics 32:671-681

Islam T, Louviere J J, Burke P F (2007) Modelling the effects of including/excluding attributes in choice experiments on systematic and random components. International Journal of Research in Marketing 24:289-300

Jaeger S R, Rose J M (2008) Stated choice experimentation, contextual influences and food choice: A case study. Food quality and preference 19:539-564

Johnson F R, Ozdemir S, Phillips K A (2010) Effects of simplifying choice tasks on estimates of taste heterogeneity in stated choice surveys. Social Science and Medicine 70:183-190

Jones L R, Cherry C R, Vu T A, Nguyen Q N (2013) The effect of incentives and technology on the adoption of electric motorcycles: A stated choice experiment in Vietnam. Transportation Research A 57:1-11

Kaenzig J, Heinzle S L, Wüstenhagen R (2013) Whatever the consumer wants, the customer gets? Exploring the gap between consumer preferences and default electricity products. Energy Policy 53:311-322

Kataria M, Bateman I, Christensen T, Dubgaard A, Hasler B, Hime S, Ladenburg J, Levin G, Martinsen L, Nissen C (2012) Scenario realism and welfare estimates in choice experiments-

A non-market valuation study on the European water framework directive. Journal of Environmental Management 94:25-33

Kehlbacher A, Balcombe K, Bennet R (2013) Stated attribute non-attendance in successive choice experiments. Journal of Agricultural Economics 64 (3):693-706

Kolstad J R (2011) How to make rural jobs more attractive to health workers. Findings from a discrete choice experiment in Tanzania. Health Economics 20:196-211

Kontoleon A, Yabe M (2003) Assessing the impacts of alternative 'opt-out' formats in choice experiments studies: Consumer preferences for genetically modified content and production information in food. Journal of Agricultural Policy research 5:1-43

Kosenius A, Ollikainen M (2013) Valuation of environmental and societal trade-offs of renewable energy sources. Energy Policy 62:1148-1156

Kragt M E, Bennet J W (2012) Attribute framing in choice experiments: How do attribute level descriptions affect value estimates? Environmental and Resource Economics 51:43-59

Lancaster K (1966) A New Approach to Consumer Theory. J Polit Econ 74:132–157

Leong W, Hensher D A (2012) Embedding multiple heuristics into choice models: An exploratory analysis. Journal of Choice Modelling 5:131-144

Lindhjem H, Navrud S (2011) Using internet in stated preference surveys: A review and comparison of survey modes. International Review of Environmental and Resource Economics 5(4):309-351

Liu P, Li Z (2012) Task complexity: A review and conceptualization framework. International Journal of Industrial Ergonomics 42:553-568

Loewenstein G, Lerner J S (2003) The role of affect in decision making. In R J Davidson, K R Scherer, H H Goldsmith. Handbook of Affective Sciences. Oxford University Press

Louviere J J (2006) What you don’t know might hurt you: some unresolved issues in the design and analysis of discrete choice experiments. Environmental and Resource Economics 34:173-188

Louviere J J, Flynn T N, Carson R T (2010) Discrete choice experiments are no tot conjoint analysis. Journal of Modelling 3 (3):57-72

Louviere J J, Islam T, Wasi N, Street D, Burgess L (2008) Designing discrete choice experiments: Do optimal designs come at a price? Journal of Consumer Research 35:360-375

Louviere JJ, Pihlens D, Carson R (2011) Design of Discrete Choice Experiments: A Discussion of Issues That Matter in Future Applied Research. J Choice Model 4:1–8

Luisetti T, Bateman I J, Turner R K (2011) Testing the fundamental assumption of choice experiments: Are values absolute or relative? Land Economics 87(2):284-296

Maddala T, Phillips K A, Johnson F R (2003) An experiment on simplifying conjoint analysis design for measuring preferences. Health Economics 12:1035-1047

Malhotra N K (1984) Reflections on the information overload paradigm in consumer decision making. Journal of Consumer Research 10 (4):436-440

Marshak J (1960) Binary choice constraints on random utility indicators. In Arrow K-Stanford Symposium on Mathematical Methods in the Social Sciences. Stanford University Press

Mazzotta M J, Opaluch J J (1995) Decision making when choices are complex: A test of Heiner’s hypothesis. Land Economics 71(4):500-515

McFadden D (1974) Conditional Logit analysis of qualitative choice behavior. In Zarembka P-Frontiers in Econometrics. Academic Press: New York

McFadden D (1978) Modeling the choice of residential location. In A. Karlquist, L. Lundquist, F. Snickars, J. W. Weibull, Spatial interaction theory and planning models. Amsterdam: North Holland, 75-96

McNair B J, Bennet J, Hensher D A (2011) A comparison of responses to single and repeated discrete choice questions. Resource and Energy Economics 33(3):554-571

Mentzakis E, Ryan M, McNamee P (2011) Using discrete choice experiments to value informal care tasks: exploring preference heterogeneity. Health Economics 20:930-944

Meyerhoff J, Liebe U (2008) Do protest responses to a contingent valuation question and choice experiment differ? Environmental and Resource Economics 39:433-446

Meyerhoff J, Liebe U (2009) Status quo effect in choice experiments: Empirical evidence on attitudes and choice task complexity. Land Economics 85 (3):515-528

Meyerhoff J, Oehlmann M, Weller P (2013) The influence of design dimension on stated choice. An example from environmental valuation using a design of design approach. Working Paper on Management in Environmental Planning 33/2013

Miller G A (1955) The magical number seven, plus or minus two. Some limits on our capacity of processing information. Psychological Review 101 (2):343–352

Morey E, Thiene M, De Salvo M, Signorello G (2008) Using attitudinal data to identify latent classes that vary in their preference for landscape preservation. Ecological Economics 68: 536-546

Mowen J C (1993) Consumer Behavior. MACMILLIAN-NEW YORK

Nguyen T C, Robinson J, Kaneko S, Komatsu S (2013) Estimating the value of economic benefits associated with adaptation to climate change in a developing country: A case study of improvements in tropical cyclone warning services. Ecological Economics 86:117-128

Ohler T, Le A, Louviere J J, Swait J (2000) Attribute range effects in binary response tasks. Marketing Letters 11(3), 249-260

Olsen S B (2009) Choosing between internet and mail surveys modes for choice experiments surveys considering non-market goods. Environmental and Resource Economics 44, 591-610

Olsen S B, Lundhede T H, Jacobsen J B, Thorsen B J, (2011) Tough and easy choices: Testing the influence of utility differences on stated certainty in choice in choice experiments. Environmental and Resource Economics 49:491-510

Pedersen L B, Gyrd-Hansen D (2013) The use of status quo and opt out options in choice experiments. Implications of researchers dubious of the neither option. International Choice Modelling Conference, Sydney-July 2013

Pedersen L B, Kjær T, Kragstrup J, Gyrd-Hansen D (2011) Does the inclusion of a cost attribute in forced an unforced choices matter? Results from a web survey applying the discrete choice experiment. Journal of Choice Modelling 4 (3):88-109

Pederesen L B, Kjær T, Kragstrup J, Gyrd-Hansen D (2012) General practitioners' preferences for the organization of primary care: A discrete choice experiment. Health Policy 106:246-25

Popkin J H, Duke J M, Borchers A M, Ilvento T (2013) Social costs from proximity to hydraulic fracturing in New York state. Energy Policy 62:62-69

Qin P, Carlsson F, Xu J (2011) Forest tenure reform in China: A choice experiment on farmers' property rights preferences. Land Economics 87:473-487

Reiter T, Völkl A, Fellendorf M (2013) Innovative approaches for an interactive stated choice survey. 92nd Annual Meeting of the Transportation Research Board TRB, Washington DC, US

Rolfe J, Bennet J (2009) The impact of offering two versus three alternatives in choice modeling experiments. Ecological Economics 68:1140-1148

Rolfe J, Bennet J, Louviere JJ (2002) Stated values and reminders of substitute goods: Testing for framing effects with choice modeling. The Australian Journal of Agricultural and Resource Economics 46 (1):1-20

Rose J M, Black I R (2006) Means matter, but variance matter too: Decomposing response latency influences on variance heterogeneity in stated preference experiments. Marketing letters 17: 295-310

Rose J M, Bliemer M C J (2004) The design of stated choice experiments: the state of practice and future challenges. ITS-WP-04-09

Rose J M, Bliemer M C J (2009) Constructing efficient choice experiments. Transport Reviews 29 (5):587-617

Rose J M, Bliemer M C J, Hensher D A, Collins A T (2008) Designing efficient stated choice experiments in the presence of reference alternatives. Transportation Research B 42:395-406

Rose J M, Hensher D A (2004) Handling individual specific availability of alternatives in stated choice experiments. In Travel survey methods: Quality and future directions-Ed.P. Stopher and C. Stecher. Elsevier, Oxford, 347-371

Rose J M, Hensher D A (2013) Demand for taxi services: new elasticity evidence. Transportation. DOI 10.1007/s11116-013-9482-5

Rose J M, Hensher D A (2013b) Toll roads are only part of the overall trip: the error of our ways in past willingness to pay studies. Transportation DOI 10.1007/s11116-013-9494-1

Rose J M, Hess S (2009) Dual-response choices in pivoted stated choice experiments. Transportation Research Record: Journal of the Transportation Research Board, No. 2135: 25-33

Rose J M, Scarpa R, Bliemer M C J (2009a) Incorporating model uncertainty into the generation of efficient stated choice experiments: A model averaging approach. ITLS-WP-09-08

Rose J M, Hensher D A, Cassuade S, de Dios Ortùzar J, Jou R C (2009b) Identifying differences in willingness to pay due to dimensionality in stated choice experiments: a cross country analysis. Journal of Transport Geography 17:21-29

Rose J M, Hess S, Bliemer M C J, Daly A (2011) The impact of varying the number of repeated choice observations on the mixed multinomial Logit model. 90th Annual Meeting of the Transportation Research Board TRB, Washington DC, US

Rosenberg R S, Peterson G L, Clarke A, Brown T C (2003) Measuring dispositions for lexicographic preferences of environmental goods: integrating economics, psychology and ethics. Ecological Economics 44:63-76

Ruby M C, Johnson F R, Mathews K E (1998) Assessing opt out options for discrete choice stated preferences: Results from a saltwater angling survey. AAEA Meeting, Salt Lake City, UT

Sándor Z, Franses H (2009) Consumer price evaluations through choice experiments. Journal of Applied Econometrics 24:517-535

Sándor Z, Wedel M (2001) Designing conjoint choice experiments using managers' prior beliefs. Journal of Marketing Research 38:430-444

Sælensminde K (2006) Causes and consequences of lexicographic choices in stated choice studies. Ecological Economics 59:331-340

Savage S J, Waldman D M (2008) Learning and fatigue during choice experiments: a comparison of online and mail survey modes. Journal of Applied Econometrics 23:351-371

Scarpa R, Willis K G, Acutt M (2007) Valuing externalities from water supply: Status quo, choice complexity and individual random effects in panel kernel Logit analysis of choice experiments. Journal of Environmental Planning and Management 50 (4):449-466

Scarpa R, Zanoli R, Bruschi V, Naspetti S (2013) Inferred and stated attribute non attendance in food choice experiments. American Journal of Agricultural Economics 95(1):165-180

Scheufele G, Bennett J (2012) Response strategies and learning in discrete choice experiments. Environmental and Resource Economics 52:435-453

Severin V (2001) Comparing statistical and respondent efficiency in choice experiments. Ph.D. Dissertation, Faculty of Economics and Business, The University of Sydney

Simon H A (1955) A behavioral model of rational choice. Quarterly Journal of Economics 69: 99-118

Sikka N, Hanley P (2013) What do commuters think travel time reliability is worth? Calculating economic value of reducing the frequency and extent of unexpected delays. Transportation 40:903-919

Sivey P, Scott A, Witt J, Joyce C, Humphreys J (2012) Junior doctors' preferences for specialty choice. Journal of Health Economics 31:813-823

Snowball J D, Willis K G (2011) Interview versus self-completion questionnaires in discrete choice experiments. Applied Economics Letters 18:1521-1525

Stopher P R, Hensher D A (2000) Are more profiles better than fewer. Searching for parsimony and relevance in stated choice experiments. Transportation Research Record 1719:165-174

Strazzera E, Mura M, Contu D (2012) Combining choice experiments with psychometric scales to assess the social acceptability of wind farm energy projects: A latent class approach. Energy Policy 48:334-347

Street D J, Burges L (2004a) Optimal and near-optimal pairs for the estimation of effects in a 2-level choice experiments. Journal of Statistical Planning and Inference 118:185-199

Street D J, Burges L (2004b) Optimal stated preference choice experiments when all choice sets contain a specific option. Statistical Methodology 1:37-45

Street D J, Burges L, Louviere J J (2005) Quick and easy choice sets: constructing optimal and nearly optimal stated choice experiments. International Journal of Research in Marketing 22: 459-470

Swait J, Adamowicz W (2001) Choice environment. Market complexity, and consumer behavior: a theoretical and empirical approach for incorporating decision complexity into models of consumer choice. Organizational behavior and human decision processes 86 (2):141-167

Swait J, Adamowicz W, Hanemann M, Diederich A, Krosnick J, Layton D, Provencher W, Schkade D, Tourangeau R (2002) Context dependance and aggregation in disaggregate choice analysis. Marketing Letters 13(3):195-205

Swait J, Louviere J J (1993) The role of the scale parameter in the estimation and comparison of Multinomial Logit models. Journal of Marketing Research 30 (3):305-314

Thurstone LL (1927) A law of comparative judgment. Psychological Review 34:273-286

Thurstone LL (1931) The indifference function. Journal of Social Psychology 2:139-167

Tudela A, Rebolledo G (2006) Optimal design of stated preference experiments when using mixed Logit models. In: Proceedings of the European Transport Conference-The Netherlands

Viney R, Norman R, Brazier J, Cronin P, King M T, Ratcliffe J, Street D (2013) An Australian discrete choice experiment to value EQ-5D health states. Health Economics DOI: 10.1002/hec.2953

Viney R, Savage E, Louviere J (2005) Empirical investigation of experimental design properties of discrete choice experiments in health care. Health Economics 14:349-362

Vista A B, Rosenberger R S, Collins A R (2009) If you provide it, will they read it? Response time effects in a choice experiment. Canadian Journal of Agricultural Economics 57:365-377

Waerden Van der P, Borgers A, Timmermans H (2004) The effects of attribute level definition on stated choice behavior. Proceedings of the 7th International Conference on Travel Survey Methods

Waerden van der P, Borgers A, Timmermans H, Bérénos M (2006) Order effects in stated-choice experiments. Study of transport mode decisions. Transportation Research Record: Journal of the Transportation Research Board No. 1985:12-18

Wardman M, Ibáñez J N (2012) The congestion multiplier: Variations in motorists' valuation of travel time with traffic conditions. Transportation Research A 46:213-225

Watson V, Sussex J, Ryan M, Tetteh E (2012) Managing poorly performing clinicians: Health care providers' willingness to pay for independent help. Health Policy 104:260-271

Wilde J, Rolfe J (2011) Comparing responses from internet and paper-based collection methods in more complex stated preference environmental valuation surveys. Economic Analysis and Policy 41(1):83-97

Wittink D, Huber J, Zandan P, Johnson R M (1992) The number of level effect in conjoint: Where does it come from and can it be eliminated? Sawtooth Software Research Paper Series

Wittink D, Krishnamurthi L, Reibstein D J (1989) The effect of differences in the number of attribute levels on conjoint results. Marketing Letters 1(2):113-123

Yu J, Goos P, Vanderbroek M (2012) A comparison of different Bayesian design criteria for setting up preference studies. Transportation Research Part B 46:789-807

Ziegler A (2012) Individual characteristics and stated preferences for alternative energy sources and propulsion technologies in vehicles: A discrete choice analysis for Germany. Transportation Research Part A 46:1372-1385

UNIONE EUROPEA
Fondo sociale europeo

REGIONE AUTONOMA DELLA SARDEGNA

FSE 2007-2013
POR
SARDEGNA
obiettivo competitività regionale e occupazione