Preference Instability in stated choice surveys: more evidence

Romain Crastes dit Sourd[*†] Stephane Hess[*‡] Maria Borejesson[*§¶]
David Hensher[‖**]

January 31, 2020

Abstract

Stated choice surveys are extensively used to inform policy making and decisions in a wide range of domains including transport, environment, health and marketing. It is common place to use multiple choice tasks per respondents in order to increase the amount of information collected. A key concern in such a context is whether the same sensitivities drive choices in all choice tasks. There are many reasons to suggest that it is not the case. Preferences could vary because of respondent fatigue, learning of preferences, lack of incentive compatibility, anchoring or other inertia effects. Many concerns have been raised about the first choice task. For some authors, it is the only one which is valid while for some others, this is merely a learning exercise. Overall, welfare estimates have been found to be stable across choice tasks. In this paper, we specifically take no position as to the validity of these claims but we note that the literature on the stability of welfare estimates has largely relied on multinomial logit models without mixing. In this paper, we propose a simple model specification for measuring whether both fixed and random coefficients are stable between the first choice task and the subsequent ones. We test the proposed specification in four different datasets and find that it yields significantly different outcomes in all four cases. We discuss the impact of our findings on data collection and stated choice modelling.

Keywords: preference stability, welfare estimates, stated preference surveys, random heterogeneity

[*]Institute for Transport Studies and Choice Modelling Centre, University of Leeds (UK)
[†]r.crastesditsourd@leeds.ac.uk
[‡]s.hess@leeds.ac.uk
[§]VTI Stockholm, Sweden
[¶]maria.borjesson@vti.se
[‖]University of Sydney Business School, Australia
[**]david.hensher@sydney.edu.au

1 Context

Stated choice (SC) is a popular survey design for studying choice behaviour. This methodology has been used for several decades across many areas of research, especially transport, marketing, health as well as environmental and resource economics (Carlsson, 2011; Hensher, 1994; Hoyos, 2010; Ryan et al., 2007). A key application of SC data and the subsequent estimation of discrete choice models is the derivation of monetary valuations, commonly referred to as willingness-to-pay (WTP) measures. These look at the monetary value that respondents place on a unit change in the characteristics of products or alternatives, and can be either for private goods (e.g. travel time) or public goods (e.g. forest preservation). While much of the emphasis is on marginal WTP, work especially in environmental economics has also looked at the WTP for an entire package or programme (Crastes et al., 2014).

The use of SC techniques has become widely accepted for producing policy guidance, but criticism has never abated completely. While all fields in which SC techniques are used have their strong believers as well as opponents of the technique, nowhere has the level of debate been as extensive as when applied to non-market valuation. Indeed, environmental economics mainly addresses the management of complex public goods, such as forest biodiversity, which has given rise to many concerns regarding whether respondents are familiar enough with the good being valued (LaRiviere et al., 2014), are not acting strategically during the survey (Scheufele and Bennett, 2012) and believe in the survey settings overall. These concerns however are again not area specific, and transport modellers have similarly discussed hypothetical and strategic bias in great detail (Hensher, 2010; Loomis, 2014).

An essential and more precise concern in SC surveys is whether the same sensitivities drive choices in all choice tasks (CTs) and to what extent is it a concern for practitioners. Indeed, a typical SC survey consists in asking respondents to complete a series of CTs where they must each time state which alternative they prefer among a finite set. Repeating the choice exercise several times allows an analyst to collect more information on respondents' preferences and the trade-offs they make. Hess et al. (2012) list three important mechanisms that are triggered by the repetition of the choice exercise: learning, fatigue and boredom. Day et al. (2012) add a distinction and separate *institutional learning* (the respondent improves his understanding of the SP rules) from *value learning* (the respondent gains knowledge about his own preferences). Börjesson and Fosgerau (2015) also found results supporting that survey participants learn about their preferences along the choice tasks and respond faster to later choices. In addition, *preference anchoring*, *framing* and issues related to the lack of *incentive compatibility* of SP (leading to strategic behaviour) have also been reported (Carson and Groves, 2007; Czajkowski et al., 2014; Ladenburg and Olsen, 2008). These effects are known as *ordering effects*.

The literature on institutional learning, fatigue and boredom has dismissed a lot of concerns about whether respondents provide less accurate responses as they start

experiencing fatigue or boredom. Hess et al. (2012) and Czajkowski et al. (2014) have both recently shown using models featuring choice task specific scale parameters that the variance in the utility function's error term can decrease (meaning that the scale of the utility function increases) as respondents complete choice tasks, contrary to early (and very influential evidence) in Bradley and Daly (1997) and Hensher and Bradley (1993). Other evidence includes Adamowicz et al. (1998), Hanley et al. (2002), Phillips et al. (2002), Hole (2004) and Holmes and Boyle (2005). It is worth noting that a few authors have reported the presence of some fatigue effect after a learning phase (Bateman et al., 2008b; Brazell and Louviere, 1996; Brouwer et al., 2010; Caussade et al., 2005; Hu, 2006). However, there are no reports of cases where fatigue outweighs institutional learning. Accommodating scale heterogeneity has little or no impact on substantive model results (Hess et al., 2012).

Far more concerns have been raised about whether the taste parameter estimates are the same for all CTs. Evidence is mixed. In a particularly important contribution on anchoring, Meyerhoff and Glenk (2015) show that even instructional choice sets, made to teach respondents how to choose before the actual SP starts, can induce anchoring effects. More precisely, respondents' preferences have been found to depend on the values they faced while being instructed about the survey settings (price levels were found to be particularly influential) which is a result also reported by Ladenburg and Olsen (2008) for female respondents only. However, the results reported by Meyerhoff and Glenk (2015) are more general in the sense that the authors systematically varied the levels of both the cost and the non-cost attributes in the instructional choice sets. The literature has also reported that respondents anchored their preferences to the first *actual* choice task they faced, especially if the good was unfamiliar (Day et al., 2012; Herriges and Shogren, 1996). It is worth noting that randomising the order in which the CTs appear in the survey for each respondent contributes to mitigate ordering effects.

The first CT has also been found to be a source of bias in the literature on *incentive compatibility*. Indeed, the seminal literature on this topic states that the message space of a choice question cannot be larger than a single binary comparison without restricting the space of allowable preference functions, that is a single binary choice is the only elicitation format that has a potential to be incentive compatible (Carson and Groves, 2007). This first suggests that respondents should not face more than two alternatives. However, Hensher (2006) as well as Meyerhoff and Liebe (2009) and Meyerhoff et al. (2015) have investigated the effects of the number of alternatives on preferences and found that the number of alternatives only has a marginal effect on model outcomes. More particularly, Hensher (2006) found that differences in behavioural responses due to the dimensionality of SC surveys exist when each dimension is assessed without controlling for the dimensions. Moreover, a strict interpretation of an *incentive compatible* design suggests that respondents only face a single CT. Indeed, since respondents may be made aware in advance of having to face multiple CTs, they can exploit information about previous CTs and decisions as they go through the survey and act strategically rather than revealing their true preferences (Czajkowski et al., 2014). This challenges the views

of Plott (1996) who have argued that respondents do not necessarily know their true preferences, and they may discover them through a preference learning process as they go through the survey.

To sum up, using multiple CTs presents both desirable and undesirable properties, and the relationship between the first CT and the subsequent ones is of particular interest. The literature on preference stability is not conclusive. The important contribution of Czajkowski et al. (2014) advocates the use of CTs specific multinomial logit models (MNL) (accounting for heterogeneity in sample level preferences only) to investigate preference variations across choice tasks. The authors do not report any differences in taste across CTs. In their review, Czajkowski et al. (2014) have shown that out of 23 studies investigating ordering effects, only 3 measured preference heterogeneity across choice tasks while accounting for random heterogeneity across respondents and none reported clear, conclusive results (Bateman et al., 2008a; Brouwer et al., 2010; Hensher, 2001). Moreover, none of these surveys investigates whether random parameters can significantly vary across CTs. This may be due to the fact that estimating CT specific models that allow for random parameters is very challenging because it is difficult to distinguish between random heterogeneity and the iid extreme value term in the Mixed Multinomial Logit Model (MMNL) (Czajkowski et al., 2014; Fosgerau and Nielsen, 2010; Hess and Train, 2011; Revelt and Train, 1998; Ruud, 1996).

In what follows, we demonstrate that a conservative approach such as using CT specific MNL models (Czajkowski et al., 2014) might lead an analyst to erroneously conclude that preferences are stable across CTs. We propose to estimate MMNL models with CT specific covariates to capture differences in fixed rand random coefficients across CTs. We argue that MNL models are unlikely to capture the shifts in parameters across CTs adequately and that the inclusion of random parameters may lead to different insights. We focus on the differences between the first CT and the subsequent ones instead of assessing whether all the CTs are different from one another. There are two reasons for this: there is a strong focus on the properties of the first CT in various field as reviewed above, and estimating MMNL models with CT specific parameters for all CTs is computationally too intensive and can lead to severe identification issues. Using four different datasets from the environment and transport literature, we estimate for each dataset a MMNL where the mean and the standard deviation of each randomly distributed parameter is different for the first CT and the subsequent ones. Moreover, a Cholesky decomposition allows to measure whether the distributions for the first CTs and the subsequent ones are correlated. Such a specification essentially allows us to measure whether there are differences between the first CT and the others without taking a position as to what drives these differences. In other words, the model specification we propose can be used as a preference stability test before conducting further investigations. Our results indicate systematic differences between the first choice task and the subsequent ones.

The remainder of this paper is organised as follows. In the next section, we outline the empirical testing framework used in our analysis. In section 3, we present the different

datasets used in our empirical work. Section 4 presents model results and finally, Section 5 concludes.

2 Method

We propose to investigate differences between the first CT and the subsequent ones using a new modelling framework followed by a series of empirical tests.

2.1 Modelling work

We build our models step by step and start by describing the well-known MMNL specification. Let U_{int} be the utility that respondent n derives from alternative i in CT t. It is made up of a modelled component V_{nit} and a random component ε_{int} which follows a type 1 extreme value distribution. We have:

$$U_{int} = V_{int} + \varepsilon_{int} \tag{1}$$

$$V_{int} = ASC_i + \beta'_n x_{int} \tag{2}$$

where β_n is a vector of taste coefficients and x_{int} a vector of attributes for alternative i. In addition, we include alternative specific constants (ASCs) for all but one of the alternatives. As a result, the probability that respondent n chooses a given alternative i conditional on β_n and the ASCs in choice situation t corresponds to the MNL probabilities

$$P_{int}(\beta_n) = \frac{e^{V_{int}}}{\sum_{j=1}^{J} e^{V_{jnt}}} \tag{3}$$

The elements in β_n can be allowed allowed to vary randomly across respondents (excluding the ASCs), using a joint distribution $f(\beta_n|\Omega)$, where Ω is a vector of parameters to be estimated, relating to the means and covariance structure of the elements in β_n. More precisely, for each one of the k elements in β_n we use the following specification:

$$\beta_{kn} = \mu 1_k + \sigma 1_k \zeta 1_{kn} \tag{4}$$

where $\mu 1_k$ corresponds to the mean and $\sigma 1_k$ the standard deviation of the random parameter. $\zeta 1_{kn}$ is a random disturbance distributed $N(0,1)$. Indeed, as the actual value of β_n for a given respondent is not observed by the analyst, the choice probabilities

are given by a multi-dimensional integral of the MNL probabilities described in Equation 5. The probability of the sequence of choices observed for person n is given by

$$L_{nt} = \int_{\beta_n} \prod_{t=1}^{T} P_{nt}(\beta_n) f(\beta_n|\Omega)\delta\beta \tag{5}$$

where P_{nt} corresponds to the probability of respondent n choosing the alternative that he was observed actually to choose. Our second model allows for differences in the means of the randomly distributed taste parameters when $CT > 1$:

$$\beta_{kn} = \mu 1_k + \sigma 1_k \zeta 1_{kn} + \mu 2_k \cdot (CT > 1) \tag{6}$$

where $\mu 2_k$ corresponds to a shift in the mean when $t > 1$ and $CT > 1$ is an indicator function which takes the value 0 for the first choice task and 1 for all others. The remainder of the model is specified the same as the base MMNL. Finally, we propose a specification which allows for differences in both fixed and random parameters between the first CT and the subsequent ones. We use:

$$\begin{aligned} &\beta_{kn} = \mu 1_k + \\ &(\sigma 1_k \zeta 1_{kn}) \cdot (CT = 1) + \\ &(\mu 2_k + \sigma 2_k \zeta 1_{kn} + \sigma 3_k \zeta 2_{kn}) \cdot (CT > 1) \end{aligned} \tag{7}$$

where $\mu 1_k$ now corresponds to the mean and $\sigma 1_k$ the standard deviation of the random parameter when $t = 1$ ($CT = 1$ is an indicator function which takes the value 1 for the first choice task and 0 else). $\zeta 1_{kn}$ is a random disturbance distributed $N(0, 1)$. Moreover, $\mu 2_k$ corresponds to a shift in the mean when $t > 1$ and $\sigma 2_k$ and $\sigma 3_k$ capture the random heterogeneity in preferences. $\zeta 2_{kn}$ is another random disturbance distributed $N(0, 1)$. This specification not only allows to capture shifts in the mean but also differences in terms of random heterogeneity between the first CT and the subsequent ones. It is worth noting that we allow the random heterogeneity in the first CT and the random heterogeneity in the subsequent ones to be correlated which is why the random disturbance $\zeta 1_{kn}$ enters the utility for both $t = 1$ and $t > 1$. Models are estimated in Willingness-To-Pay Space (WTPS). For each one of the datasets presented in the next section, we estimate five models:

- Model A: MNL model
- Model B: shifted MNL model
- Model C: MMNL model
- Model D: Shifted MMNL model
- Model E: Shifted MMNL model with CT specific random heterogeneity

Table 1: Model specifications

Model	Specification	Choice task
Model A	$\beta_k = \mu 1_k$	All
Model B	$\beta_k = \mu 1_k$	$CT1$
	$\beta_k = \mu 1_k + \mu 2_k$	$CT > 1$
Model C	$\beta_k = \mu 1_k + \sigma 1_k \zeta 1_{kn}$	All
Model D	$\beta_k = \mu 1_k + \sigma 1_k \zeta 1_{kn}$	$CT1$
	$\beta_k = \mu 1_k + \mu 2_k + \sigma 1_k \zeta 1_{kn}$	$CT > 1$
Model E	$\beta_k = \mu 1_k + \sigma 1_k \zeta 1_{kn}$	$CT1$
	$\beta_k = \mu 1_k + \mu 2_k + \sigma 2_k \zeta 1_k + \sigma 3_k \zeta 2_{kn}$	$CT > 1$

2.2 Framework for empirical tests

2.2.1 Model fit impacts

We first compare whether allowing for different sensitivities across CTs improves model fit by comparing model A to model B, model C to model D and E and model D to model E using likelihood ratio tests. LL_A corresponds to the log-likelihood at convergence for model A and the same notation applies to the other models. We compute $-2(LL_B - LL_A) \sim \chi^2_{R-1}$, $-2(LL_D - LL_C) \sim \chi^2_{R-1}$, $-2(LL_E - LL_C) \sim \chi^2_{R-1}$ and $-2(LL_E - LL_D) \sim \chi^2_{R-1}$ where R corresponds to the number of parameters for each model.

2.2.2 Welfare estimates

Secondly, we investigate whether the mean and the standard deviation of the welfare estimates are the same for the first CT and the subsequent ones depending on whether random heterogeneity is considered or not. More precisely, we compute WTP or Value of Travel Time (VTT) estimates for each model specification and group of CT (first one and subsequent ones) and investigate whether:

1. different model specifications lead to differences in the mean. Mean welfare estimates are compared within and across models using T-tests.

2. the difference between the first CT and the other ones for each attribute differs depending on the specification used.

2.2.3 Differences in distributions

Our last test consists in plotting the kernel density estimate of each of the WTP (or VTT) distributions derived from the shifted MMNL model with CT specific random heterogeneity. Kernel density estimation (KDE) is simply a non-parametric technique

for estimating the probability density function of a random variable. We then use the k-density test for comparing the common area of KDE proposed by Martínez-Camblor et al. (2008). This test allows to assess how similar or different two distributions are. More precisely, the k-density test gives a simple measure of the proximity of two kernel density estimates. This measure, known as the AC statistic, varies between 0 and 1. A value of 0 corresponds to an absolute discordance while a value of 1 corresponds to an absolute match of the distributions. For each dataset, we test whether and by how much the AC statistic differs between the first CT and the subsequent ones for the shifted MMNL model with CT specific random heterogeneity.

3 Data

We use four SP surveys datasets from different countries (Australia, Denmark and Poland). The SP surveys vary in terms of design (number of attributes, number of choice scenarios, and number of alternatives). Overall, we use two datasets from non-market valuation surveys and two datasets from transport surveys. By including data from such a diverse set of surveys, we can establish whether differences exist across areas. For each survey, the order of the CTs was randomised across individual participants. Figure 1 reports one choice task for each one of the datasets featured in our analysis.

3.1 Danish value of time survey

Our first case study makes use of data from a choice survey conducted in Denmark (more details can be found in Fosgerau (2006)). In each of the 8 choice tasks, a respondent was faced with two unlabelled alternatives described by travel time (tt) and travel cost (tc).

3.2 Sydney toll road dataset

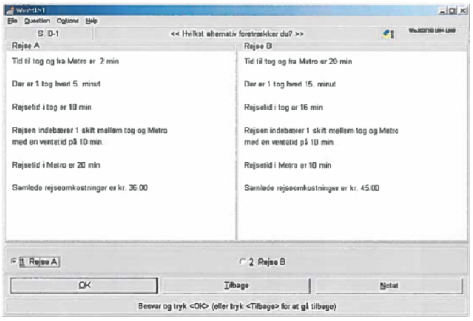
Our second case study makes use of data from a three alternative route choice experiment in Australia (one alternative consisted of a reference trip and was kept fixed across choice tasks). The alternatives were described in term of free flow time (ff), slowed down time (sdt), running costs ($cost$), tolls ($toll$) and travel time variability (var). More details can be found in Hensher and Rose (2005).

3.3 Bialowieza Forest survey

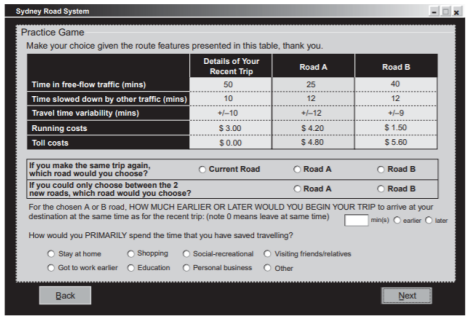
The third case study used data presented by Bartczak (2015), among others. The survey features 12 choices tasks using 3 alternatives each and using attributes describing changes in the quality of the Bialowieza forest (Poland). The choice experiment comprised 4 attributes: cen (level of naturalness of the commercial part of the forest), gos (level of naturalness of the second-growth forest), $vis1$ and $vis2$ (restrictions on number of visitors per day) and fee (annual cost per household).

3.4 Ecological value of Polish forests survey

The fourth case study used data collected by Czajkowski et al. (2012). The survey consisted in eliciting the preferences of the general public in Poland for increasing the amount of recreational infrastructures in Polish forests and protecting biodiversity. Respondents faced 26 choice tasks between four alternatives. Each alternative was described by 4 attributes taking several levels: $nat1$ and $nat2$ (partial and substantial improvement in the protection of the forests), $tra1$ and $tra2$ (partial and substantial reduction of the amount of litter found in the forests) and $inf1$ and $inf2$ (partial and substantial improvements in tourist infrastructure). A price attribute was also included.



(a) Danish Value of Time survey



(b) Sydney toll road

	Program A No additional protection measures	Program B Additional protection measures	Program C Additional protection measures
National Park and Natural Reserves (35% of the Białowieża Forest)	**HIGH** level of naturalness		
Commercial forest (50% of the Białowieża Forest)	**LOW** level of naturalness	**HIGH** level of naturalness	**HIGH** level of naturalness
Second-growth forest (15% of the Białowieża Forest)	**LOW** level of naturalness	**LOW** level of naturalness	**HIGH** level of naturalness
Restriction on number of visitors	**LACK** of constraints	**MAX 7,500** people per day	**MAX 5,000** people per day
Annual cost per household	**0zł**	**50zł**	**100zł**
Your choice	□	□	□

(c) Bialowieza Forest survey

	Alternative 1	Alternative 2	Alternative 3	Alternative 4
Protection of ecologically valuable forests	**Status quo** Passive protection of **50%** of the most ecologically valuable forests (1.5% of all forests)	**Status quo** Passive protection of **50%** of the most ecologically valuable forests (1.5% of all forests)	**Status quo** Passive protection of **50%** of the most ecologically valuable forests (1.5% of all forests)	**Substantial improvement** Passive protection of 100% of the most ecologically valuable forests (3% of all forests, 100% increase)
Litter in forests	**Status quo** No change in the amount of litter in the forests	**Partial improvement** Decrease the amount of litter in the forests by half (50% reduction)	**Status quo** No change in the amount of litter in the forests	**Partial improvement** Decrease the amount of litter in the forests by half (50% reduction)
Infrastructure	**Status quo** No change in tourist infrastructure	**Status quo** No change in tourist infrastructure	**Partial improvement** Appropriate tourist infrastructure in 50% more forests (50% increase)	**Substantial improvement** Appropriate tourist infrastructure available in double the current forests (100% increase)
Cost	0 PLN	10 PLN	25 PLN	100 PLN
Your choice	□	□	□	□

(d) Ecological value of Polish forests survey

Figure 1: Choice tasks - Illustration

4 Results

In this section, we report the results obtained with each of the four datasets described above. It is worth noting that the time attribute in the Danish dataset is specified as log-uniform, while the non-monetary attributes in the Sydney dataset are all log-normally

distributed. The non-monetary attributes are all normally distributed for the non-market valuation datasets. In all cases, the cost coefficient has been specified as negative log-normal. For the fixed-coefficient models, the non-monetary $\mu 1$ parameters need to be interpreted as value of time (or willingness-to-pay) measures while for the MMNL models, they need to be interpreted as the mean of the logarithm of the corresponding VTT measures for the transport data, and as the mean of the normally distributed WTP distribution for the non-market valuation data. The other parameters must be interpreted in a similar fashion.

4.1 Danish value of time survey

We first look at the results from the Danish value of time survey. Model results are reported in Table 2 while the LR test results are reported in Table 3. We find an interesting pattern of results. Model B is found to be an improvement when compared to the simple MNL model (Model A) based on the results of an LR test. Moreover, Model D and Model E are both found to be better than Model C. Finally, Model E outperforms Model D.

Table 2: Danish value of time survey - Model results (in WTP space - 100s or Ore)

		Model A		**Model B**		**Model C**		**Model D**		**Model E**	
LL(final)		-1836.722		-1832.829		-1609.946		-1598.855		-1593.469	
Adj.Rho-square		0.0466		0.0476		0.163		0.168		0.169	
AIC		3679.44		3675.66		3229.89		3211.71		3208.94	
BIC		3697.24		3705.32		3259.55		3253.23		3274.19	
		Est.	R. T	Est.	R. T	Est.	R. T	Est.	R. T	Est.	R. T
$\mu 1$	asc	0.1049	3.12	0.1045	3.10	0.1780	3.69	0.1712	4.03	0.1662	3.33
	VTT	1.0178	11.5	1.1100	3.88	-0.6360	-6.17	1.3770	5.67	1.7662	3.39
	cost	-0.0842	-7.19	-0.0414	-2.19	-1.1278	-8.07	-5.7490	-9.70	-13.8916	-1.11
$\sigma 1$	VTT	.	.	.	.	1.1985	11.91	-3.9417	-15.18	-4.7143	-4.65
	cost	.	.	.	.	2.2585	8.61	6.5356	8.84	20.0421	0.96
$\mu 2$	VTT	.	.	-0.1039	-0.36	.	.	-0.0959	-0.38	-0.4941	-0.93
	cost	.	.	-0.0521	-2.38	.	.	1.6582	3.5279	11.7345	0.92
$\sigma 2$	VTT	.	.	.	.	.	.	.	.	-3.8609	-15.72
	cost	.	.	.	.	.	.	.	.	5.9309	4.65
$\sigma 3$	VTT	.	.	.	.	.	.	.	.	0.0000	NA
	cost	.	.	.	.	.	.	.	.	-3.3240	-1.30

Moving on to the analysis of the welfare estimates (Table 4), we note that the first choice task for Model E yields higher mean VTT estimates than the subsequent ones (the medians are found to be the same while the standard deviation of VTT is much higher for the first choice task). Model D and Model E yield higher mean VTT values in comparison to Model C. However, we note that introducing CT specific random parameters does not influence welfare measures for this dataset. Indeed, the VTT derived from the different models have not been found to be significantly different across models and choice tasks

Table 3: Danish value of time survey - Likelihood ratio test results

	LR test value	**Degrees of freedom**	**p-value**
Model B vs Model A	7.63	1	*0.005*
Model D vs Model C	22.18	2	*0.000*
Model E vs Model C	32.85	5	*0.000*
Model E vs Model D	10.67	3	*0.013*

according to a series of T-tests reported in Table 5, which should be interpreted as follows: the first result, -0.31, corresponds to the T-value related to the difference between the mean VTT for model A and the mean VTT for model B. The result is not significant at the usual levels. The remainder of the Table should be interpreted the same way.

Table 4: Danish value of time survey - Welfare estimates (100s or Ore)

Model		$ALLCT$			$CT1$			$CT>1$		
		Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Model A	VTT	1.02	.	.	.	.	.	.	.	.
Model B		1.01	.	.	1.01	.	.	1.01	.	.
Model C		0.89	0.49	0.92	.	.	.	.	.	.
Model D		0.91	0.51	0.93	0.99	0.55	1.01	0.89	0.50	0.92
Model E		0.96	0.52	1.00	1.22	0.55	1.45	0.92	0.52	0.94

Table 5: Danish value of time survey - VTT comparisons across models using T-tests

	A	B	C	$D-CT=1$	$D-CT>1$	$E-CT=1$	*Model*
VTT	-0.31	.	.	.	.	.	B
	1.14	-0.74	.	.	.	.	C
	0.08	-0.25	0.23	.	.	.	$D-CT=1$
	1.32	-0.75	0.06	0.06	.	.	$D-CT>1$
	-0.85	0.32	1.39	0.51	1.42	.	$E-CT=1$
	0.99	-0.70	0.13	0.19	0.12	-1.33	$E-CT>1$

Finally, we analyse differences in the distribution of the VTT for the first CT and the subsequent ones between Models C, D and E. Figure 2 suggests that the distribution of the VTT for the first CT for Model E appears to be slightly different that the other distributions considered, which is confirmed by a series of k-density tests. The AC statistic between the kernel density of the distribution of the VTT for the first CT for Model E and the VTT for the subsequent ones is for the same model is found to be only 0.848. Moreover, the common area of the VTT between Model E (CT1) and the other distributions is systematically below 90%, while the common area between the other distributions considered is 97% on average (meaning that the distributions are nearly identical). Altogether, these results show that there are notable differences between the

first CT and the subsequent ones for the Danish value of time data, especially when parameters are random.

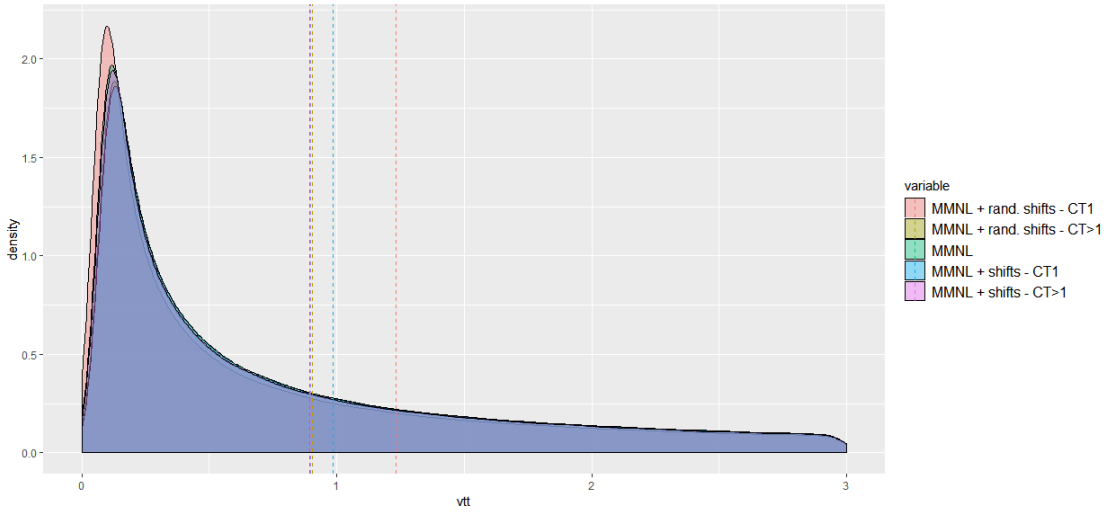


Figure 2: Danish value of time survey - VTT distributions - MMNL + rand. shifts model

Table 6: Danish value of time survey - Common Area of kernel densities

Model	Model E - $CT1$	Model E - $CT > 1$	Model D - $CT1$	Model D - $CT > 1$	
VTT	1	.	.	.	Model E - CT1
	0.848	1	.	.	Model E - $CT > 1$
	0.893	0.961	1	.	Model D - $CT1$
	0.881	0.976	0.997	1	Model D - $CT > 1$
	0.863	0.982	0.979	0.992	Model C

4.2 Sydney toll road survey

We now describe the results obtained with the second dataset, the Sydney toll road survey. Model results are reported in Table 7. The distribution of VTT_ff, VTT_sdt and VTT_var is positive log-normal while the distribution of *cost* and *toll* is negative log-normal. Models are estimated in cost space. Interestingly, we find that Model B is not an improvement over the basic MNL model (Model A) and that none of the shifts in the means introduced in Model B are significant, while Model D and Model E are both substantially improving the goodness-of-fit with respect to Model C. Model E outperforms Model D. This is confirmed below by a series of likelihood-ratio tests (see Table 8). We note that the only shift in the mean of the random parameters which has been found to be significant is for the toll attribute. As a result, all the other shifts have

been fixed to zero for Model D and Model E. We find that $\sigma 3\_VTT\_sdt$, $\sigma 3\_VTT\_var$ and $\sigma 3\_VTT\_toll$ are significant and that the standard deviations for the first CT ($\sigma 1$) are very different than the standard deviations for the subsequent CT ($\sigma 2$) for some of the parameters.

Table 7: Sydney survey - Model results (in WTP space)

		Model A		**Model B**		**Model C**		**Model D**		**Model E**	
LL(final)		-3027.662		-3024.601		-2428.145		-2421.981		-2409.58	
Adj.Rho-square		0.2895		0.2891		0.4287		0.433		0.431	
AIC		6069.32		6073.2		4880.29		4869.96		4861.16	
BIC		6113.18		6148.39		4955.48		4951.41		4992.74	
		Est.	R. T	Est.	R. T	Est.	R. T	Est.	R. T	Est.	R. T
$\mu 1$	asc1	-0.1979	-1.37	-0.1987	-1.35	-1.4025	-6.13	-1.3976	-5.02	-1.4125	-6.22
	asc2	-0.2744	-1.85	-0.2779	-1.84	-1.4860	-6.61	-1.4921	-5.28	-1.4982	-6.72
	VTT_ff	13.1598	12.30	11.3721	3.28	2.1667	12.06	2.2026	8.77	2.2366	15.18
	VTT_sdt	17.3959	20.23	12.5915	4.55	2.6247	23.58	2.6399	18.19	2.5965	28.02
	VTT_var	1.1085	1.08	1.7449	1.43	1.5134	5.21	1.5269	1.98	1.4783	5.48
	cost	-0.3135	-17.15	-0.3806	-4.23	-0.6526	-7.41	-0.6523	-7.16	-0.6355	-7.34
	toll	-0.3614	-13.87	-0.2920	-5.50	-0.6532	-7.52	-1.1145	-5.65	-1.3262	-2.44
$\sigma 1$	VTT_ff	.	.	.	.	1.1031	11.76	1.0002	14.03	0.3295	0.69
	VTT_sdt	.	.	.	.	0.6877	5.50	0.6763	2.02	0.4345	2.61
	VTT_var	.	.	.	.	1.2879	10.83	1.3032	10.60	1.0547	3.57
	cost	.	.	.	.	-0.5990	-6.99	-0.6226	-1.18	-0.6690	-4.98
	toll	.	.	.	.	-0.8452	-9.07	-0.8645	-6.56	-1.2033	-1.00
$\mu 2$	VTT_ff	.	.	2.0240	0.53	.	.	0.0000	NA	0.0000	NA
	VTT_sdt	.	.	5.3077	1.74	.	.	0.0000	NA	0.0000	NA
	VTT_var	.	.	-0.6606	-0.67	.	.	0.0000	NA	0.0000	NA
	cost	.	.	0.0731	0.80	.	.	0.0000	NA	0.0000	NA
	toll	.	.	-0.0750	-1.47	.	.	0.4984	3.08	0.7416	1.3700
$\sigma 2$	VTT_ff	.	.	.	.	.	.	.	.	1.0617	17.52
	VTT_sdt	.	.	.	.	.	.	.	.	0.5457	10.55
	VTT_var	.	.	.	.	.	.	.	.	1.4128	7.16
	cost	.	.	.	.	.	.	.	.	-0.6067	-11.91
	toll	.	.	.	.	.	.	.	.	-0.7774	-9.44
$\sigma 3$	VTT_ff	.	.	.	.	.	.	.	.	0.0000	NA
	VTT_sdt	.	.	.	.	.	.	.	.	0.4968	10.09
	VTT_var	.	.	.	.	.	.	.	.	0.4177	4.63
	cost	.	.	.	.	.	.	.	.	0.0000	NA
	toll	.	.	.	.	.	.	.	.	-0.4256	-6.85

The analysis of the welfare estimates (see Table 9) reveals that the first CT yields much lower VTT than the subsequent ones for Model E, and that these values are also much lower than those derived from Model C and D. The mean VTT values are not all found to be significantly different across models and CT according to the T-tests reported in Table 10, where significant different are reported in bold. We find significant differences between the first choice tasks and the subsequent ones for model E, where VTT_ff and VTT_var are found to be significantly lower for the first choice task, which challenges the idea that preferences are stable across CTs.

Table 8: Sydney survey - Likelihood ratio test results

	LR test value	**Degrees of freedom**	**p-value**
Model B vs Model A	6.12	5	*0.295*
Model D vs Model C	12.33	1	*0.000*
Model E vs Model D	37.13	9	*0.000*
Model E vs Model D	24.8	8	*0.001*

Table 9: Sydney survey - Welfare estimates (In AUS per hours)

Model	Attribute	*ALL*			*CT*1			*CT* > 1		
		Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Model A	*VTT_ff*	13.16	.	.	.	.	.	.	.	.
	VTT_sdt	17.40	.	.	.	.	.	.	.	.
	VTT_var	1.11	.	.	.	.	.	.	.	.
Model B	*VTT_ff*	13.38	.	.	13.16	.	.	13.40	.	.
	VTT_sdt	17.86	.	.	17.40	.	.	17.90	.	.
	VTT_var	1.09	.	.	1.11	.	.	1.08	.	.
Model C	*VTT_ff*	16.11	8.77	24.53	.		.	.		.
	VTT_sdt	17.49	13.79	13.61	.		.	.		.
	VTT_var	10.58	4.56	22.12	.		.	.		.
Model D	*VTT_ff*	14.98	9.09	19.53	.		.	.		.
	VTT_sdt	17.62	14.00	13.42	.		.	.		.
	VTT_var	10.94	4.62	23.44	.		.	.		.
Model E	*VTT_ff*	16.10	9.40	22.35	9.89	9.38	3.35	16.51	9.41	23.64
	VTT_sdt	17.47	13.44	14.42	14.75	13.41	6.71	17.65	13.41	15.07
	VTT_var	12.83	4.40	33.92	7.73	4.40	11.23	13.17	4.38	35.52

Table 10: Sydney toll road survey - VTT comparisons across models using T-tests

	C	*D*	$E - CT = 1$	$E - CT > 1$	*Model*
VTT_ff	-1.30	-0.51	**1.66**	-1.57	*A*
	.	0.295	**2.411**	-0.154	*C*
	.	.	1.377	-0.410	*D*
	.	.	.	**-2.689**	$E - CT = 1$
VTT_sdt	-0.04	-0.08	1.35	-0.13	*A*
	.	-0.041	1.105	-0.059	*C*
	.	.	0.901	-0.001	*D*
	.	.	.	-1.260	$E - CT = 1$
VTT_var	**-4.28**	-1.33	**-3.31**	**-4.87**	*A*
	.	-0.048	1.081	-0.878	*C*
	.	.	0.423	-0.296	*D*
	.	.	.	**-1.916**	$E - CT = 1$

We now look at Figure 3 and Table 11 and find that the distribution of VTT are very different between the first CT for Model E and the other distributions considered. Indeed, the common area between the distribution of VTT_ff for Model E ($CT1$) and Model E ($CT > 1$) is only 62% while it is 82% between Model C and Model E ($CT > 1$).

Similar results are observed for VTT_sdt and VTT_var[1].

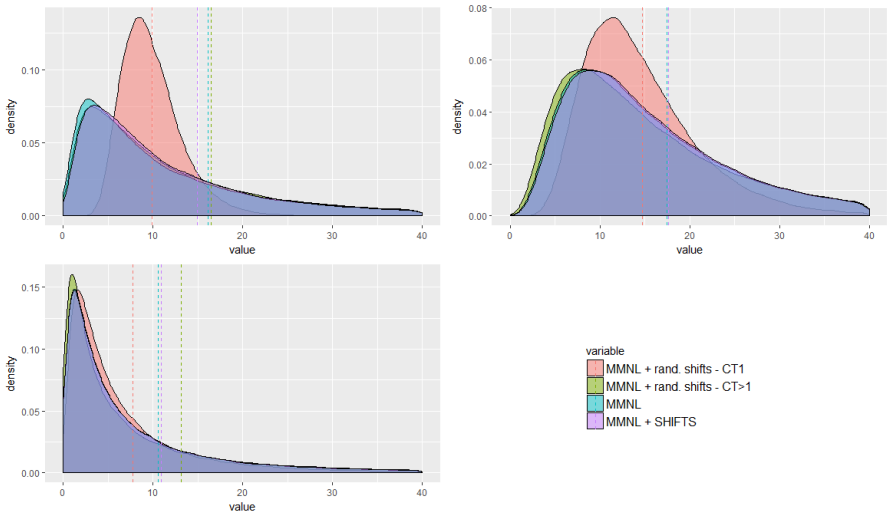


Figure 3: Sydney survey - VTT distributions - MMNL + rand. shifts model

Table 11: Sydney survey - Common area of kernel density estimates

	Model E - $CT1$	Model E - $CT > 1$	
VTT_ff	1	.	Model E - $CT1$
	0.62	1	Model E - $CT > 1$
	0.64	0.82	Model C
VTT_sdt	1	.	Model E - $CT1$
	0.57	1	Model E - $CT > 1$
	0.58	0.9	Model C
VTT_var	1	.	Model E - $CT1$
	0.69	1	Model E - $CT > 1$
	0.66	0.8	Model C

[1]Results for Model D are not reported because the distributions have been found to be very similar to Model C and potentially not informative given that the only difference between the models is the introduction of μ_toll in Model D.

4.3 Bialowieza Forest survey

Our third dataset is the Bialowieza Forest survey. Again, we observe significant differences across CTs once introducing different parameters depending on CTs (see Table 12). WTP_cen, WTP_gos, WTP_vis1 and WTP_vis2 have been specified as normally distributed, while fee is negative log-normal. We find once again that Model B does not improve model fit in comparison to Model A (Likelihood-ratio test results are provided in Table 13). In fact, none of the observed shifts in the mean are found to be significant in any of the models estimated, which led us to fix these parameters to 0 for Model D (which is hence equivalent to Model C). Looking at the results for Model E, we find that the standards deviations of the randomly distributed parameters for the first CT ($\sigma 1$) are very different from the standard deviations for the subsequent CTs ($\sigma 2$) and that most of the parameters in $\sigma 3$ are found to be significant.

Table 12: Bialowieza Forest survey - Model results

		Model A		**Model B**		**Model C**		**Model D**	**Model E**	
LL(final)		-14843.38		-14840.57		-11110.01			-11087.54	
Adj.Rho-square		0.0629		0.0627		0.2981			0.2987	
AIC		29700.76		29705.13		22244.03			22225.09	
BIC		29753.8		29796.05		22334.95			22414.5	
		Est.	R. T	Est.	R. T	Est.	R. T		Est.	R. T
$\mu 1$	asc1	-0.2219	-3.41	-0.2218	-3.40	1.4978	17.89		1.5314	20.01
	asc2	-0.3166	-4.96	-0.3165	-4.93	1.3441	16.03		1.3747	18.09
	WTP_cen	0.4498	13.95	0.4393	8.18	0.2121	21.82		0.2796	17.16
	WTP_gos	0.4171	12.73	0.3955	7.52	0.1258	13.34		0.0608	6.87
	WTP_vis1	0.0291	1.25	0.1139	1.76	-0.1175	-9.04		-0.1238	-9.01
	WTP_vis2	0.1239	4.96	0.0609	0.88	-0.0466	-3.64		-0.1783	-8.29
	fee	-1.2623	-18.34	-1.4517	-9.97	1.3807	26.93		1.8746	12.73
$\sigma 1$	WTP_cen	.	.	.	.	0.50	25.74		0.6706	28.81
	WTP_gos	.	.	.	.	0.49	16.76		0.8337	34.90
	WTP_vis1	.	.	.	.	0.16	10.57		0.2688	38.75
	WTP_vis2	.	.	.	.	0.35	12.98		0.4145	14.14
	fee	.	.	.	.	1.36	22.58	**NA**	2.1542	6.43
$\mu 2$	WTP_cen	.	.	0.0117	0.23	.	.		-0.0969	-3.60
	WTP_gos	.	.	0.0236	0.47	.	.		0.0669	4.13
	WTP_vis1	.	.	-0.0938	-1.37	.	.		0.0000	NA
	WTP_vis2	.	.	0.0692	0.97	.	.		0.1133	3.96
	fee	.	.	0.2056	1.43	.	.		-0.4940	-3.35
$\sigma 2$	WTP_cen	.	.	.	.	.	.		0.5005	29.68
	WTP_gos	.	.	.	.	.	.		0.4634	20.22
	WTP_vis1	.	.	.	.	.	.		0.1109	9.50
	WTP_vis2	.	.	.	.	.	.		0.3134	20.90
	fee	.	.	.	.	.	.		1.3039	19.49
$\sigma 3$	WTP_cen	.	.	.	.	.	.		0.0000	NA
	WTP_gos	.	.	.	.	.	.		-0.1841	-11.40
	WTP_vis1	.	.	.	.	.	.		-0.1186	-10.10
	WTP_vis2	.	.	.	.	.	.		-0.0884	-2.15
	fee	.	.	.	.	.	.		0.5437	5.41

Table 13: Bialowieza Forest survey - Likelihood ratio test results

	LR test value	**Degrees of freedom**	**p-value**
Model B *vs* Model A	5.62	5	*0.345*
Model E *vs* Model C	44.94	13	*0.000*

The results from the LR test indicate that Model E outperforms Model C. Indeed, the Likelihood ratio test-value is equal to 44.94 and the Likelihood ratio test p-value is equal to 0.000 (13 degrees of freedom). On the other hand, Model B does not outperform Model A, which is an excepted result given that all the shifts introduced in Model B were not found to be significant.

Table 14 shows that the WTP values derived from Model E ($CT1$) are very different from the values derived from Model E ($CT > 1$) and Model C, excepted for WTP_vis1 (the standard deviations are different though). Again, there is no specific pattern in the differences. WTP_cen is higher for the first CT while WTP_gos and WTP_vis2 are much lower.

The k-density tests provided below show again that the distributions are different.

Table 14: Bialowieza Forest survey - Welfare estimates (In 100s of zlotys)

Model	Attribute	*ALL*			*CT*1			*CT* > 1		
		Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Model A	WTP_cen	0.45	.	.	.	.	.	.	.	.
	WTP_gos	0.42	.	.	.	.	.	.	.	.
	WTP_vis1	0.03	.	.	.	.	.	.	.	.
	WTP_vis2	0.12	.	.	.	.	.	.	.	.
Model B	WTP_cen	0.45	.	.	0.44	.	.	0.45	.	.
	WTP_gos	0.42	.	.	0.40	.	.	0.42	.	.
	WTP_vis1	0.03	.	.	0.11	.	.	0.02	.	.
	WTP_vis2	0.12	.	.	0.06	.	.	0.13	.	.
Model C	WTP_cen	0.21	0.21	0.50	.	.	.	.	.	.
	WTP_gos	0.13	0.13	0.49	.	.	.	.	.	.
	WTP_vis1	-0.12	-0.12	0.17	.	.	.	.	.	.
	WTP_vis2	-0.05	-0.05	0.35	.	.	.	.	.	.
Model E	WTP_cen	0.19	0.19	0.51	0.28	0.28	0.67	0.18	0.18	0.50
	WTP_gos	0.12	0.12	0.52	0.06	0.06	0.83	0.13	0.13	0.50
	WTP_vis1	-0.12	-0.12	0.17	-0.12	-0.12	0.27	-0.12	-0.12	0.16
	WTP_vis2	-0.07	-0.07	0.33	-0.18	-0.18	0.42	-0.07	-0.07	0.33

Tables 16 and 15 as well as Figure 4 confirm that the distribution of WTP values derived from Model E ($CT1$) are different than the distributions derived from Model E ($CT > 1$) and Model C. The common area of kernel density estimates between Model E ($CT1$) and Model E ($CT > 1$) ranges from 0.78 to 0.89 and from 0.80 to 0.91 when

Table 15: Bialowieza Forest survey - WTP comparisons across models using T-tests

	C	$E-CT=1$	$E-CT>1$	*Model*
WTP_cen	**7.058**	**4.711**	**6.990**	A
	.	**-3.558**	1.295	C
	.	.	**-3.600**	$E-CT=1$
WTP_gos	**8.544**	**10.498**	**8.123**	A
	.	**5.026**	-0.113	C
	.	.	**-4.130**	$E-CT=1$
WTP_vis1	**5.498**	**5.656**	**5.656**	A
	.	0.333	0.333	C
	.	.	0.000	$E-CT=1$
WTP_vis2	**6.074**	**9.168**	**6.309**	A
	.	**5.262**	0.881	C
	.	.	**-3.960**	$E-CT=1$

comparing Model E ($CT1$) and Model C. On the other hand, the WTP distributions derived from Model E ($CT>1$) and Model C are found to be nearly identical.

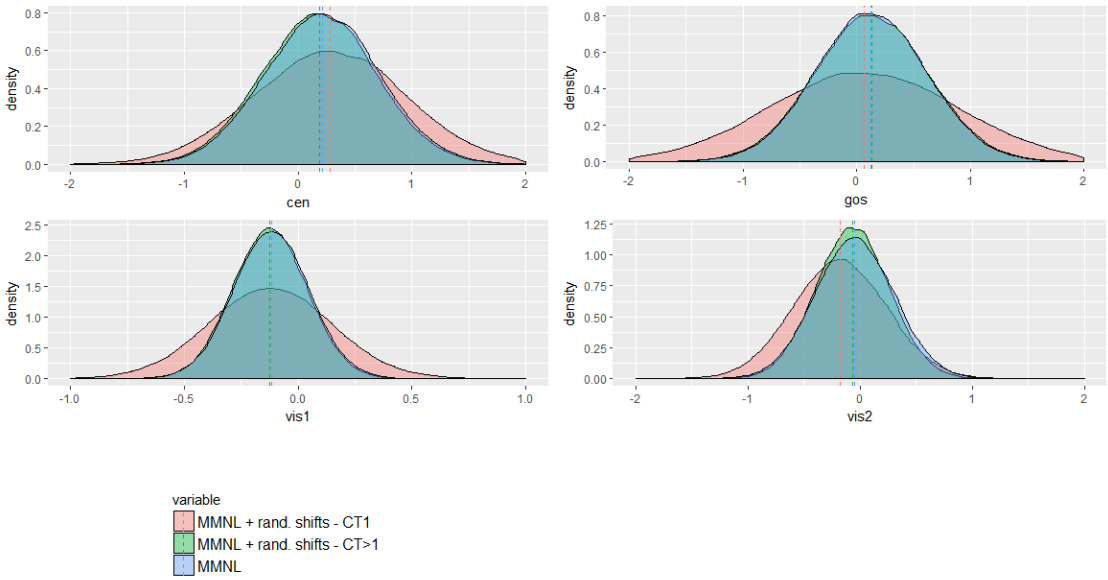


Figure 4: Bialowieza Forest survey - WTP distributions - Model E

Table 16: Bialowieza Forest survey - Common area of kernel density estimates

Model	Model E - $CT1$	Model E - $CT > 1$	
WTP_cen	1 0.886 0.912	. 1 0.995	Model E - $CT1$ Model E - $CT > 1$ Model C
WTP_gos	1 0.78 0.798	. 1 1	Model E - $CT1$ Model E - $CT > 1$ Model C
WTP_vis1	1 0.818 0.837	. 1 1	Model E - $CT1$ Model E - $CT > 1$ Model C
WTP_vis2	1 0.88 0.906	. 1 0.999	Model E - $CT1$ Model E - $CT > 1$ Model C

4.4 Ecological value of Polish forests survey

The model results for the fourth (and final dataset) are described in Table 17. WTP_nat1, WTP_nat2, WTP_tra1, WTP_tra2, WTP_inf1 and WTP_inf2 have been specified as normally distributed while fee has been specified as log-normally distributed. Some of the parameters introduced in Model B are found to be significant, while none of the shifts introduced in Model D are, meaning that Model D is not different from Model C for this dataset. For Model E, $\mu 2\_WTP\_nat1$, $\mu 2\_WTP\_nat2$, $\sigma 3\_WTP\_tra1$, $\sigma 3\_WTP\_tra2$ and $\sigma 3\_WTP\_inf1$ have been found to be significant. Again, Model B is found to be an improvement over Model A and Model E is found to be an improvement over Model C as confirmed by the LR tests reported in Table 18.

Table 17: Ecological value of Polish forests survey - Model results (in 100s of zlotys)

		Model A		**Model B**		**Model C**		**Model D**	**Model E**	
LL(final)		-29694.84		-29689.67		-21389.54			-21325.21	
Adj.Rho-square		0.1767		0.1768		0.4067			0.4081	
AIC		59409.68		59403.34		42813.08			42710.41	
BIC		59491.35		59501.34		42951.91			42955.42	
		Est.	R. T	Est.	R. T	Est.	R. T		Est.	R. T
$\mu 1$	asc1	-2.0057	-23.03	-2.0063	-22.87	-2.5468	-28.7		-2.5590	-28.88
	asc2	-1.9530	-22.42	-1.9538	-22.25	-2.4649	-28.1		-2.4762	-28.28
	asc3	-2.0545	-23.78	-2.0555	-23.61	-2.6119	-29.45		-2.6234	-29.59
	WTP_nat1	0.5935	16.45	0.4367	6.62	0.8195	19.93		0.6291	7.55
	WTP_nat2	0.8732	16.78	0.6936	9.29	1.1525	20.18		1.0304	9.47
	WTP_tra1	1.0672	18.23	1.0685	18.77	1.1549	19.68		1.0752	12.09
	WTP_tra2	1.4275	18.60	1.4297	19.27	1.6868	19.98		1.6985	17.94
	WTP_inf1	0.4862	17.12	0.4872	17.54	0.5366	19.96		0.5310	19.47
	WTP_inf2	0.7827	18.76	0.7834	19.36	0.8485	21.6		0.8522	19.05
	fee	-1.3442	-20.73	-1.3419	-21.69	0.4192	7.12		0.4171	5.53
$\sigma 1$	WTP_nat1	.	.	.	.	-0.51	-15.32		-0.5282	-5.79
	WTP_nat2	.	.	.	.	0.72	13.54		0.5470	5.24
	WTP_tra1	.	.	.	.	-0.12	-1.73		-0.3395	-2.89
	WTP_tra2					-0.44	-13.85		-0.4205	-4.38
	WTP_inf1					-0.29	-12.07		-0.8404	-6.11
	WTP_inf2	.	.	.	.	-0.33	-12.81		-0.1996	-1.94
	fee					1.30	25.41	**NA**	1.2432	11.58
$\mu 2$	WTP_nat1	.	.	0.1647	2.68	.	.		0.2302	3.20
	WTP_nat2	.	.	0.1885	3.11	.	.		0.1762	2.05
	WTP_tra1	.	.	0.0000	NA	.	.		0.0802	1.20
	WTP_tra2			0.0000	NA				0.0000	NA
	WTP_inf1			0.0000	NA				0.0000	NA
	WTP_inf2	.	.	0.0000	NA	.	.		0.0000	NA
	fee			0.0000	NA				0.0000	NA
$\sigma 2$	WTP_nat1	.	.	.	.	.	.		-0.4931	-15.10
	WTP_nat2	.	.	.	.	.	.		0.6720	13.85
	WTP_tra1	.	.	.	.	.	.		-0.1263	-5.73
	WTP_tra2								-0.4142	-13.33
	WTP_inf1	.	.	.	.	.	.		-0.1294	-4.34
	WTP_inf2								-0.3450	-12.27
	fee								1.3442	26.89
$\boldsymbol{\sigma 3}$	WTP_nat1	.	.	.	.	.	.		0.0000	NA
	WTP_nat2	.	.	.	.	.	.		0.0000	NA
	WTP_tra1	.	.	.	.	.	.		0.0901	3.06
	WTP_tra2								0.1759	5.72
	WTP_inf1	.	.	.	.	.	.		-0.2459	-8.04
	WTP_inf2								0.0000	NA
	fee								0.0000	NA

Table 18: Ecological value of Polish forests survey - Likelihood ratio test results

	LR test value	**Degrees of freedom**	**p-value**
Model B vs Model A	10.34	2	*0.006*
Model E vs Model D	128.66	13	*0.000*

Tables 19 and 20 (T-tests) indicate that the mean WTP values derived from Model E ($CT1$) are different than those derived from Model E ($CT > 1$) and Model C for WTP_nat1, and the mean WTP for $nat1$ and $nat2$ are different between Model E ($CT = 1$) and Model E ($CT > 1$). The WTP estimates are found to be higher for the models with mixing (C and E) compared to the models without (A and B).

Table 19: Ecological value of Polish forests survey - Welfare estimates (In 100s of zlotys)

Model	Attribute	*ALL*			*CT*1			*CT* > 1		
		Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Model A	WTP_nat1	0.59	.	.	.	.	.	.	.	.
	WTP_nat2	0.87	.	.	.	.	.	.	.	.
	WTP_tra1	1.07	.	.	.	.	.	.	.	.
	WTP_tra2	1.43	.	.	.	.	.	.	.	.
	WTP_inf1	0.49	.	.	.	.	.	.	.	.
	WTP_inf2	0.78	.	.	.	.	.	.	.	.
Model B	WTP_nat1	0.60	.	.	0.44	.	.	0.60	.	.
	WTP_nat2	0.87	.	.	0.69	.	.	0.88	.	.
	WTP_tra1	1.07	.	.	1.07	.	.	1.07	.	.
	WTP_tra2	1.43	.	.	1.43	.	.	1.43	.	.
	WTP_inf1	0.49	.	.	0.49	.	.	0.49	.	.
	WTP_inf2	0.78	.	.	0.78		.	0.78	.	.
Model C	WTP_nat1	0.82	0.82	0.51	.	.	.	.	.	.
	WTP_nat2	1.15	1.15	0.72	.	.	.	.	.	.
	WTP_tra1	1.15	1.15	0.12	.	.	.	.	.	.
	WTP_tra2	1.69	1.69	0.44	.	.	.	.	.	.
	WTP_inf1	0.54	0.54	0.29	.	.	.	.	.	.
	WTP_inf2	0.85	0.85	0.33	.	.	.	.	.	.
Model E	WTP_nat1	0.85	0.85	0.49	0.63	0.63	0.53	0.86	0.86	0.49
	WTP_nat2	1.20	1.20	0.66	1.03	1.03	0.54	1.21	1.21	0.67
	WTP_tra1	1.15	1.15	0.16	1.07	1.07	0.34	1.15	1.15	0.16
	WTP_tra2	1.70	1.70	0.45	1.70	1.70	0.42	1.70	1.70	0.45
	WTP_inf1	0.53	0.53	0.28	0.53	0.53	0.84	0.53	0.53	0.28
	WTP_inf2	0.85	0.85	0.34	0.85	0.85	0.20	0.85	0.85	0.35

Finally, the results from Table 21 and Figure 5 indicate that the distribution of WTP derived from Model E ($CT1$) are different than the distributions derived from Model E ($CT > 1$) and Model C, excepted for WTP_tra2. This difference is very large for

Table 20: Ecological value of Polish forests survey - WTP comparisons across models using T-tests

	B	$B2$	C	$E-CT=1$	$E-CT>1$	*Model*
	2.085	-0.156	**-4.131**	-0.392	**-4.048**	A
	.	**-2.680**	**-4.925**	**-1.810**	**-4.925**	B
WTP_nat1	.	.	**-4.020**	-0.306	**-3.950**	$B2$
	.	.	.	**2.049**	-0.580	C
	.	.	.	.	**3.200**	$E-CT=1$
	1.973	-0.123	**-3.615**	-1.303	**-3.827**	A
	.	**-3.110**	**-4.882**	**-2.552**	**-5.017**	B
WTP_nat2	.	.	**-3.544**	-1.236	**-3.762**	$B2$
	.	.	.	0.994	-0.599	C
	.	.	.	.	**2.050**	$E-CT=1$
	-0.016	-0.016	-1.058	-0.075	-0.981	A
	.	0.000	-1.057	-0.063	-0.978	B
WTP_tra1	.	.	-1.057	-0.063	-0.978	$B2$
	.	.	.	0.748	-0.006	C
	.	.	.	.	1.200	$E-CT=1$
	-0.021	-0.021	**-2.273**	**-2.224**	**-2.224**	A
	.	0.000	**-2.288**	**-2.235**	**-2.235**	B
WTP_tra2	.	.	**-2.288**	**-2.235**	**-2.235**	$B2$
	.	.	.	-0.092	-0.092	C
	.	.	.	.	0.000	$E-CT=1$
	-0.025	-0.025	-1.289	-1.138	-1.138	A
	.	0.000	-1.278	-1.125	-1.125	B
WTP_inf1	.	.	-1.278	-1.125	-1.125	$B2$
	.	.	.	0.146	0.146	C
	.	.	.	.	0.000	$E-CT=1$
	-0.012	-0.012	-1.148	-1.136	-1.136	A
	.	0.000	-1.154	-1.141	-1.141	B
WTP_inf2	.	.	-1.154	-1.141	-1.141	$B2$
	.	.	.	-0.062	-0.062	C
	.	.	.	.	0.000	$E-CT=1$

WTP_tra1 since the common area measure is found to be only 0.65 between Model E ($CT1$) and Model E ($CT>1$) (0.55 for Model D).

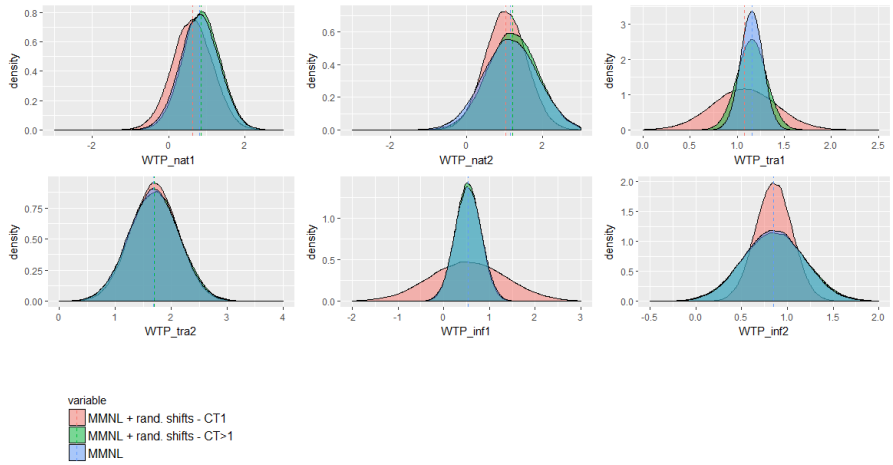


Figure 5: Ecological value of Polish forests survey - WTP distributions - MMNL + rand. shifts model

Table 21: Ecological value of Polish forests survey - Common area of kernel density estimates

Model	**Model E -** $CT1$	**Model E -** $CT > 1$	
WTP_nat1	1 0.871 0.9	. 1 0.995	Model E - $CT1$ Model E - $CT > 1$ Model C
WTP_nat2	1 0.892 0.884	. 1 0.981	Model E - $CT1$ Model E - $CT > 1$ Model C
WTP_tra1	1 0.656 0.551	. 1 0.895	Model E - $CT1$ Model E - $CT > 1$ Model C
WTP_tra2	1 0.993 0.998	. 1 1	Model E - $CT1$ Model E - $CT > 1$ Model C
WTP_inf1	1 0.546 0.561	. 1 1	Model E - $CT1$ Model E - $CT > 1$ Model C
WTP_inf2	1 0.772 0.788	. 1 1	Model E - $CT1$ Model E - $CT > 1$ Model C

5 Discussion and conclusions

This research is an attempt to provide a reliable although limited answer to the ongoing debate on whether the preferences expressed by respondents in a repeated SC survey format, which is by far the most widely used format in a large stream of fields including transport, environment and marketing, are stable across CT. We have noted that most of the surveys who have investigated preference stability have used MNL models with fixed coefficients and CT specific shifts, or CT specific MNL models (which are rigorously the same) (Czajkowski et al., 2014; Hess et al., 2012). We have hypothesised that models which feature random parameters, such as the MMNL, have not been widely used for measuring preference instability. We have also noted that the literature has mostly expressed concerns about the first CT when it comes to parameters stability. As a result, we have proposed a new, simple model specification based on the MMNL model where we specify different random parameters for the first CT and the subsequent ones. We tested the proposed model specification on four different datasets from both Transport research and Non-market Valuation. Our results are summarised in Table 22 below (X indicates whether significant differences between parameter estimates across CTs have been found or not).

Table 22: Results - Summary

	Model B	Model D	Model E	
Case study	$\mu 2$	$\mu 2$	$\mu 2$	$\sigma 3$
Danish value of time survey	X	X	X	X
Sydney toll road survey	.	X	X	X
Bialowieza Forest survey	.	.	X	X
Ecological value of Polish forests survey	X	.	X	X

We found conclusive evidences relating to preference instability. Our findings indicate that models that feature choice task specific random parameters outperform the models that do not in all the four cases. Little evidence has been found for supporting the stability of preferences when using repeated SC surveys. More importantly, we found that each one of the four datasets considered exhibited a different pattern of outcomes. Results for the Danish value of time survey showed that adding CT specific random parameters fits the data better but does not yield significantly different mean VTTs, while results for the Sydney toll road survey indicated that the fixed parameter model (Model B) did not report any differences, in contrary to the random parameter models (Model D and E), which did. Moreover, the Bialowieza forest survey models only reported significant results when allowing for both shifts in the mean and the standard deviation of the random parameters, while for the Ecological Value of Polish forests survey it has been found that Model B and Model E provided significant results but not Model D.

The use of several statistical tests including likelihood ratio tests, T-tests and k-density tests for measuring the common area of kernel density estimates show that models which feature CT specific random parameters fit the data better (as proven by likelihood ratio tests), provide different welfare estimates in comparison to a classic MMNL model and, depending on whether the WTP (or VTT) parameters are normally or log-normally distributed, we find that mean welfare measures can be different for the first CT and the subsequent ones for the same model.

Overall, our results clearly indicate that preferences are not stable between the first choice tasks and the subsequent ones and that simple MNL models might not be necessarily able to reveal such results. Moreover, our results also indicate a clear absence of pattern in terms of whether the first CT yields higher (or lower) welfare measures in comparison to the subsequent ones. At the very least, the results we obtained should lead to question which values and CT should be considered for welfare analysis.

6 Acknowledgements

The authors acknowledge the financial support by the European Research Council through the consolidator grant 615596-DECISIONS. The authors would also like to thank Mikołaj Czajkowski for his valuable contributions to this work.

References

Adamowicz, W., Boxall, P., Williams, M., and Louviere, J. (1998). Stated preference approaches for measuring passive use values: choice experiments and contingent valuation. *American journal of agricultural economics*, 80(1):64–75.

Bartczak, A. (2015). The role of social and environmental attitudes in non-market valuation: an application to the białowieża forest. *Forest Policy and Economics*, 50:357–365.

Bateman, I., Carson, R. T., Day, B., Dupont, D., Louviere, J. J., Morimoto, S., Scarpa, R., Wang, P., et al. (2008a). Choice set awareness and ordering effects in discrete choice experiments.

Bateman, I. J., Carson, R. T., Day, B., Dupont, D., Louviere, J. J., Morimoto, S., Scarpa, R., and Wang, P. (2008b). Choice set awareness and ordering effects in discrete choice experiments in discrete choice experiments. Technical report, CSERGE working paper EDM.

Bradley, M. A. and Daly, A. J. (1997). Estimation of logit choice models using mixed stated preference and revealed preference information. *Understanding travel behaviour in an era of change*, pages 209–232.

Brazell, J. and Louviere, J. (1996). Helping, learning, and fatigue: An empirical investigation of length effects in conjoint choice studies. *Department of Marketing, The University of Sydney*.

Börjesson, M. and Fosgerau, M. (2015). Response time patterns in a stated choice experiment. *Journal of Choice Modelling*, 14:48 – 58.

Brouwer, R., Dekker, T., Rolfe, J., and Windle, J. (2010). Choice certainty and consistency in repeated choice experiments. *Environmental and Resource Economics*, 46(1):93–109.

Carlsson, F. (2011). Non-market valuation: stated preference methods. *The Oxford handbook of the economics of food consumption and policy*, 181:214.

Carson, R. T. and Groves, T. (2007). Incentive and informational properties of preference questions. *Environmental and resource economics*, 37(1):181–210.

Caussade, S., de Dios Ortúzar, J., Rizzi, L. I., and Hensher, D. A. (2005). Assessing the influence of design dimensions on stated choice experiment estimates. *Transportation research part B: Methodological*, 39(7):621–640.

Crastes, R., Beaumais, O., Arkoun, O., Laroutis, D., Mahieu, P.-A., Rulleau, B., Hassani-Taibi, S., Barbu, V. S., and Gaillard, D. (2014). Erosive runoff events in the european union: Using discrete choice experiment to assess the benefits of integrated management policies when preferences are heterogeneous. *Ecological economics*, 102:105–112.

Czajkowski, M., Giergiczny, M., and Greene, W. H. (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*.

Czajkowski, M., Giergiczny, M., and Greene, W. H. (2014). Learning and fatigue effects revisited: Investigating the effects of accounting for unobservable preference and scale heterogeneity. *Land Economics*, 90(2):324–351.

Day, B., Bateman, I. J., Carson, R. T., Dupont, D., Louviere, J. J., Morimoto, S., Scarpa, R., and Wang, P. (2012). Ordering effects and choice set awareness in repeat-response stated preference studies. *Journal of environmental economics and management*, 63(1):73–91.

Fosgerau, M. (2006). Investigating the distribution of the value of travel time savings. *Transportation Research Part B: Methodological*, 40(8):688–707.

Fosgerau, M. and Nielsen, S. F. (2010). Deconvoluting preferences and errors: A model for binomial panel data. *Econometric Theory*, 26(6):1846–1854.

Hanley, N., Wright, R. E., and Koop, G. (2002). Modelling recreation demand using choice experiments: climbing in scotland. *Environmental and resource Economics*, 22(3):449–466.

Hensher, D. A. (1994). Stated preference analysis of travel choices: the state of practice. *Transportation*, 21(2):107–133.

Hensher, D. A. (2001). Measurement of the valuation of travel time savings. *Journal of Transport Economics and Policy (JTEP)*, 35(1):71–98.

Hensher, D. A. (2006). Towards a practical method to establish comparable values of travel time savings from stated choice experiments with differing design dimensions. *Transportation Research Part A: Policy and Practice*, 40(10):829–840.

Hensher, D. A. (2010). Hypothetical bias, choice experiments and willingness to pay. *Transportation Research Part B: Methodological*, 44(6):735–752.

Hensher, D. A. and Bradley, M. (1993). Using stated response choice data to enrich revealed preference discrete choice models. *Marketing Letters*, 4(2):139–151.

Hensher, D. A. and Rose, J. M. (2005). Respondent behavior in discrete choice modeling with a focus on the valuation of travel time savings. *Journal of Transportation and Statistics*, 8(2):17.

Herriges, J. A. and Shogren, J. F. (1996). Starting point bias in dichotomous choice valuation with follow-up questioning. *Journal of environmental economics and management*, 30(1):112–131.

Hess, S., Hensher, D. A., and Daly, A. (2012). Not bored yet–revisiting respondent fatigue in stated choice experiments. *Transportation research part A: policy and practice*, 46(3):626–644.

Hess, S. and Train, K. E. (2011). Recovery of inter-and intra-personal heterogeneity using mixed logit models. *Transportation Research Part B: Methodological*, 45(7):973–990.

Hole, A. R. (2004). Forecasting the demand for an employee park and ride service using commuters' stated choices. *Transport Policy*, 11(4):355–362.

Holmes, T. P. and Boyle, K. J. (2005). Dynamic learning and context-dependence in sequential, attribute-based, stated-preference valuation questions. *Land Economics*, 81(1):114–126.

Hoyos, D. (2010). The state of the art of environmental valuation with discrete choice experiments. *Ecological economics*, 69(8):1595–1603.

Hu, W. (2006). Effects of endogenous task complexity and the endowed bundle on stated choice. Technical report.

Ladenburg, J. and Olsen, S. B. (2008). Gender-specific starting point bias in choice experiments: Evidence from an empirical study. *Journal of Environmental Economics and Management*, 56(3):275–285.

LaRiviere, J., Czajkowski, M., Hanley, N., Aanesen, M., Falk-Petersen, J., and Tinch, D. (2014). The value of familiarity: effects of knowledge and objective signals on willingness to pay for a public good. *Journal of Environmental Economics and Management*, 68(2):376–389.

Loomis, J. B. (2014). 2013 waea keynote address: Strategies for overcoming hypothetical bias in stated preference surveys. *Journal of Agricultural and Resource Economics*, pages 34–46.

Martínez-Camblor, P., De Una-Alvarez, J., and Corral, N. (2008). k-sample test based on the common area of kernel density estimators. *Journal of Statistical Planning and Inference*, 138(12):4006–4020.

Meyerhoff, J. and Glenk, K. (2015). Learning how to choose—effects of instructional choice sets in discrete choice experiments. *Resource and Energy Economics*, 41:122–142.

Meyerhoff, J. and 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., and Weller, P. (2015). The influence of design dimensions on stated choices in an environmental context. *Environmental and resource economics*, 61(3):385–407.

Phillips, K. A., Johnson, F. R., and Maddala, T. (2002). Measuring what people value: a comparison of "attitude" and "preference" surveys. *Health Services Research*, 37(6):1659–1679.

Plott, C. (1996). *The Rational Foundations of Economic Behaviour*, chapter Rational Individual Behavior in Markets and Social Choice Processes: the Discovered Preference Hypothesis, pages 225–250. McMillian , London.

Revelt, D. and Train, K. (1998). Mixed logit with repeated choices: households' choices of appliance efficiency level. *Review of economics and statistics*, 80(4):647–657.

Ruud, P. (1996). Approximation and simulation of the multinomial probit model: an analysis of covariance matrix estimation. *Department of Economics, Berkeley*, pages 1–17.

Ryan, M., Gerard, K., and Amaya-Amaya, M. (2007). *Using discrete choice experiments to value health and health care*, volume 11. Springer Science & Business Media.

Scheufele, G. and Bennett, J. (2012). Response strategies and learning in discrete choice experiments. *Environmental and Resource Economics*, 52(3):435–453.