Preference instability in stated choice surveys: more evidence (abridged version)

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

January 31, 2020

Wordcount: 2220 words + 6 Tables + 1 Figure

1 Context

Stated choice (SC) is a popular survey design for studying choice behaviour, especially in Transport research. This methodology has been used for several decades across many other areas of research, including marketing, health as well as environmental and resource economics (Carlsson, 2011; Hensher, 1994). A key application of SC data and the subsequent estimation of discrete choice models is the derivation of monetary valuations, commonly referred to as Value of Travel Time (VTT) or 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.

An essential issue in SC surveys is whether the same sensitivities drive choices in all choice tasks (CTs) and to what extent is it something practitioners should worry about. 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. Specific concerns have been raised about whether the taste parameter estimates are the same for all CTs. However, most of the research work carried out about this topic has relied on CT specific multinomial logit models (MNL) (Czajkowski et al., 2014), where taste heterogeneity is not specified as random across the population.

*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

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 fields, 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 transport and environment literature[1], 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.

2 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

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

[1]We only report results for one dataset in this abridged version. The full paper features results from the four dataset mentioned.

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). 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	$\beta_k = \mu 1_k$ $\beta_k = \mu 1_k + \mu 2_k$	$CT1$ $CT > 1$
Model C	$\beta_k = \mu 1_k + \sigma 1_k \zeta 1_{kn}$	All
Model D	$\beta_k = \mu 1_k + \sigma 1_k \zeta 1_{kn}$ $\beta_k = \mu 1_k + \mu 2_k + \sigma 1_k \zeta 1_{kn}$	$CT1$ $CT > 1$
Model E	$\beta_k = \mu 1_k + \sigma 1_k \zeta 1_{kn}$ $\beta_k = \mu 1_k + \mu 2_k + \sigma 2_k \zeta 1_k + \sigma 3_k \zeta 2_{kn}$	$CT1$ $CT > 1$

3 Framework for empirical tests

3.1 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.

3.2 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.

3.3 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.

4 Results

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 transport surveys and two datasets from non-market valuation 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. In what follows, we only describe results for one of the datasets.

The case study reported in this extended abstract 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). Model results are compared in Table 2 and details outputs are reported in Table 3. 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. 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 $\sigma3\_VTT\_sdt$, $\sigma3\_VTT\_var$ and $\sigma3\_VTT\_toll$ are significant and that the standard deviations for the first CT ($\sigma1$) are very different than the standard deviations for the subsequent CT ($\sigma2$) for some of the parameters.

Table 2: 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*

The analysis of the welfare estimates (see Table 4) 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 5, 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. Finally, we look at Figure 1 and Table 6 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[2].

[2]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.

Table 3: 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

Table 4: 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 5: 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**	$E - CT = 1$

Table 6: Sydney survey - Common area of kernel density estimates

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

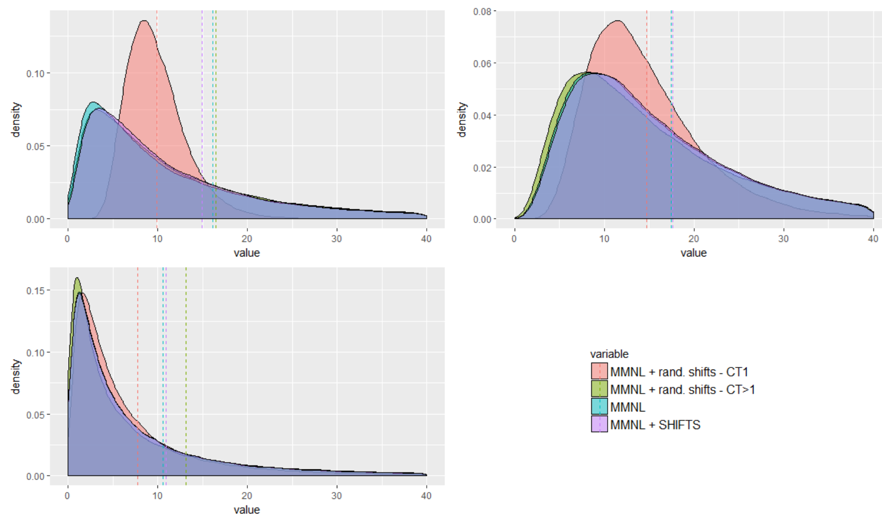

Figure 1: Sydney survey - VTT distributions - MMNL + rand. shifts model

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).

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. As previously mentioned, the final paper features results from four different dataset and is ready for presentation.

References

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

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.

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

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.

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.

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.