TI 2013-074/VIII
Tinbergen Institute Discussion Paper

Changing with the Tide: Semi-Parametric Estimation of Preference Dynamics

Thijs Dekker[1]
Paul Koster[2]
Roy Brouwer[3]

[1] Section of Transport and Logistics, Delft University of Technology;
[2] Department of Spatial Economics, VU University Amsterdam, Tinbergen Institute;
[3] Department of Environmental Economics, Institute for Environmental Studies, VU University Amsterdam.

Changing with the tide: semi-parametric estimation of preference dynamics

THIJS DEKKER[a], PAUL KOSTER[b] and ROY BROUWER[c]

[a] *Delft University of Technology, Section of Transport and Logistics, Jaffalaan 5, 2628BX Delft, the Netherlands. (e-mail: t.dekker@tudelft.nl; corresponding author)*
[b] *Department of Spatial Economics & Tinbergen Institute Amsterdam, VU University Amsterdam, the Netherlands. (e-mail: p.r.koster@vu.nl)*
[c] *Department of Environmental Economics, Institute for Environmental Studies, VU University Amsterdam, the Netherlands. (e-mail: r.brouwer@vu.nl)* [1]

Abstract

This paper contrasts the discovered preference hypothesis against the theory of coherent arbitrariness in a split-sample stated choice experiment on flood risk exposure in the Netherlands. A semi-parametric local multinomial logit model (L-MNL) is developed as an alternative to the Swait and Louviere (1993) procedure to control for preference dynamics within and between samples. The L-MNL model finds empirical support for the discovered preference hypothesis in the form of a declining starting point bias induced by the first choice task. These results differ from the Swait and Louviere procedure which, due to its limited flexibility, accepts the standard assumption underlying microeconomic theory of stable preference parameters throughout the choice sequence. The observed preference dynamics puts the use of choice experiments at risk of generating biased welfare estimates if not controlled for.

Keywords: Preference dynamics; Discovered preference hypothesis; Coherent arbitrariness; Preference uncertainty; Local multinomial logit model

JEL classification numbers – C14, D12, Q51, Q54

[1] Acknowledgements: This research was funded by the Dutch National Research Program 'Climate Changes Spatial Planning' (www.klimaatvoorruimte.nl). Also the financial support from the ERC (Advanced Grant OPTION #246969) for Paul Koster's contribution is gratefully acknowledged. The authors would like to thank Hans Koster for providing useful comments on an earlier version of this paper. We also would like to thank participants of the EAERE Conference in Rome 2010 for insightful discussions.

I. Introduction

Non-market valuation studies strongly rely on discrete choice models to quantify the welfare implications of changes in the provision of goods and services as measured through Willingness-To-Pay (WTP) and Consumer Surplus. In particular, stated choice experiments repeatedly present respondents with comparable choice tasks in order to improve the statistical efficiency of discrete choice models for a given sample size. Holmes and Boyle (2005) point out the risk of obtaining biased welfare estimates as a result of not or incorrectly controlling for preference dynamics over the choice sequence.

This paper contrasts two competing hypotheses regarding preference dynamics, namely the Discovered Preference Hypothesis (Plott, 1996) and the theory of Coherent Arbitrariness (Ariely et al. 2003). Both hypotheses deviate from the standard microeconomic assumption of well-defined and stable preferences underlying the random utility maximization model (McFadden, 2001). The Discovered Preference Hypothesis (DPH) assumes well-defined preferences exist before respondents come to the (hypothetical) market. However, initially these preferences are not fully known to the respondent. Through repetition and market experience, individuals discover and learn about their 'true' preferences. Under the DPH hypothesizes preferences converge to this underlying set of well-defined preferences making the convergence level path independent (Bateman et al. 2008; Braga and Starmer 2005). In contrast, the preference construction literature argues that a stable set of preferences is non-existent prior to a stated choice survey (e.g. Ariely et al. 2003; Ariely et al. 2006; Fischhoff et al. 1999; McFadden 1999; Slovic 1995). Ariely et al. (2003) stipulate in their theory of Coherent Arbitrariness (CA) that individuals gradually develop a set of stable preferences due to an internal drive for consistency. Consistency with past decisions drives future choices, making the convergence process path dependent. Consequently, (arbitrary) initial value cues are expected to influence the level at which preference converge.

Closely related to the existence and stability of well-defined preferences is the discussion regarding the existence and decay of a starting point bias (SPB). The latter has been extensively discussed in the contingent valuation literature and more recently also in the choice experiment literature (Carlsson and Martinsson 2008; Groeneveld 2010; Ladenburg and Olsen 2008). This paper applies the concept of a SPB to induce alternative starting points across two independent subsamples. The resulting preference dynamics and related convergence levels are examined to empirically contrast the DPH and the theory of CA.[2]

The second aim of this paper is to improve the empirical identification of preference dynamics in discrete choice models. The paper offers two contributions to the literature. First, an improved experimental design enables better identification of dynamics in welfare measures over the choice sequence. Second, a novel econometric approach is developed, named the local multinomial logit (L-MNL) model, as an alternative to the commonly applied Swait and Louviere (1993) test procedure to control for preference dynamics (e.g. Bech et al. 2011; Brouwer et al. 2010; Carlsson et al. 2010; Holmes and Boyle 2005; Ladenburg and Olsen 2008). The Swait and Louviere (1993) test is subject to the risk of under- and over-smoothing of preference dynamics, because it either treats the preference relation in two alternative choice tasks either as identical or as independent. The latter introduces a trade-off between bias and efficiency. The L-MNL model provides an intermediate solution by controlling the degree of smoothing (Fröhlich 2006, Koster

[2] Inducing a SPB is one approach to contrast the competing hypotheses comprised in the DPH and theory of CA.The presence of a starting point bias is interpreted as if a stable set of preferences is not known to the respondent at the start of the survey. A decay of the starting point bias supports the DPH, while a persistent starting point bias supports the theory of CA. Support for either hypothesis remains conditional on a stable set of preferences by the end of the choice sequence. Alternative tests contrasting both hypotheses may, for example, alter the order of the choice sequence, thereby inducing path dependence.

and Koster 2013). By estimating choice task specific models it prevents bias and over-smoothing effects. By also drawing information from closely related observations, i.e. related in terms of their position in the choice task and sample membership, efficiency issues and under-smoothing are dealt with.

Based on the responses to a stated choice experiment on flood risk exposure in the Netherlands, specifically designed to contrast the DPH and the theory of CA, we reach different conclusions using the L-MNL model, a set of independent models, and the Swait and Louviere (1993) test. The Swait and Louviere (1993) test supports the micro-economic framework by finding limited preference dynamics. Its limited flexibility, however, forces the method to account for subtle preference dynamics through the scale parameter. On the contrary, the set of independent models finds erratic patterns of preference dynamics questioning their overall stability. The L-MNL model, as an intermediate, but flexible model form, provides a more consistent picture of within and between sample preference dynamics. Specifically, it supports the DPH by finding a gradual decay of a starting point bias. The observed preference dynamics are mainly related to the tendency to select the Status Quo option, but marginal WTP estimates also converge between the samples. At the start of the choice sequence, respondents are less likely to select the Status Quo, but become less willing to make trade-offs between the attributes included in the choice experiment as the choice sequence progresses. The initial willingness to trade-off can be amplified by presenting high price levels in the first choice task.

The structure of the paper is as follows. Section II describes the context, underlying hypotheses and experimental set-up of the paper. Section III discusses the properties of the L-MNL model. Section IV provides a description of the case study and Section V covers the analytical results. Section VI concludes.

II. Empirical approach: contrasting the DPH and the theory of CA

Within and between sample preference dynamics

This study controls for two types of preference dynamics. First, it tests for *within-sample* preference dynamics over the choice sequence, where both the DPH and CA predict the emergence of a stable set of preferences due to learning. Second, it tests for *between-sample* preference dynamics at the choice task level. The DPH predicts convergence in preferences across samples, whereas the theory of CA predicts convergence towards a set of stable, but sample specific preferences subject to arbitrary initial value clues.

In a controlled experimental setting two independent samples are presented with sample specific initial value cues. The two samples are referred to respectively as the Low Starting Bid (LSB) and High Starting Bid (HSB) sample. The only difference in the experimental set-up between the LSB and HSB samples arises in the first choice task.[3] Both samples are presented with an identical initial choice card containing exactly the same alternatives, attributes and associated attribute levels. Only the levels of the cost attribute differ across the two samples. The LSB sample is assigned the lowest levels of the price vector and the HSB sample the highest levels. In the remaining choice tasks, both samples are presented with choice cards from exactly the same experimental design. Hence, this study only takes into account starting point effects, but not the effect of showing different attribute levels to different respondents (e.g. Carlsson and Martinsson 2008; Hanley et al. 2005; Morkbak et al. 2010; Ohler et al. 2000).

By presenting respondents with either a high or a low value cue in the first choice task, differences in cost sensitivity across samples are expected. Respondents in the LSB sample have a lower reference value induced by the experimental design. In subsequent choice tasks they are

[3] The words choice task and choice card are used in this paper. The latter refers to a specific choice situation as included in the experimental design. The former refers to the position of the choice card in the choice sequence.

presented with alternatives associated with higher (or comparable) costs, possibly making them less willing to make trade-offs, i.e. having a higher cost sensitivity. The opposite effect is expected for the HSB sample, because in this sample respondents are subsequently presented with cheaper alternatives. Their lower cost sensitivity makes them more willing to make trade-offs across attributes. Therefore, higher marginal WTP estimates are expected in the HSB sample compared to the LSB sample, reflecting a SPB. As respondents proceed through the remaining choice tasks, they encounter different attribute levels and learn about their preferences. Consequently, the impact of the initial choice task on subsequent choices is expected to decay. In accordance with the DPH, marginal WTP estimates are expected to stabilize and converge between both samples.

Improving the experimental design

Testing for *within* and *between* sample preference dynamics requires the estimation of choice task specific preference parameters within the LSB and HSB samples. Estimation of a choice model for a specific choice task in each sample requires that all choice cards in the experimental design are answered a sufficient number of times at each moment during the choice sequence. If this is not the case, parameter estimates will have high standard errors, because only a limited number of trade-offs are considered. Identifying whether preference dynamics are the result of 'true' preference dynamics, limitations of the design, or heterogeneity in preferences across respondents becomes hard under these circumstances. This may have played a role, for example, in Ladenburg and Olsen (2008), where each respondent was presented with the same choice task at the same moment in the choice sequence. This paper reduces the limitations of the design by applying a rotating procedure, where the order in which the choice cards are presented is structurally varied across respondents.

The experimental design is identical for the HSB and LSB sample and consists of three blocks of eight choice cards each.[4] The total set of 24 choice cards was generated in Ngene (NGENE, 2010). The three blocks of eight choice cards are used in the rotating procedure. That is, version 1 presents respondents with block one (choice cards 1-8 in ascending order). Version 2 starts with choice cards 2-8 and ends with choice card 1. Similarly, version 9 presents block two (cards 9-16 in ascending order), while version 10 starts with choice cards 10-16 and ends with choice card 9. This rotation procedure yields 24 versions in total. Finally, the order of appearance of the first and second unlabeled policy alternatives on each choice card between which respondents are asked to choose is altered to prevent a reading bias from left to right. Accordingly, the number of versions doubles to 48 and respondents are randomly assigned to one version. As a result of the rotation procedure, on average, each choice card in the design is answered on average ten times at each moment in the design by respondents from a particular sample. More details about the experimental design and response frequencies are found in Appendix A.

This careful experimental set-up minimizes the possibility that within and between sample preference dynamics identified during the analysis can be attributed to limitations of the experimental design. The design only affects the first choice task in order to induce an SPB. Furthermore, to minimize the impact of heterogeneity in respondent preferences on choice task specific parameter estimates, respondents for both samples are obtained independently from a representative sample of panel participants. Demographic and socio-economic characteristics are

[4] We do not include responses of the first choice task in our analysis. Accordingly, we will estimate sixteen specific models, one for each of the eight choice tasks within each sample separately.

monitored during the survey and are therefore expected to be the same at the sample level. Accordingly, preference dynamics identified during the analysis are assumed to contain less noise.

III. Econometric Methods to test for preference dynamics over the choice sequence

The Swait and Louviere (1993) test procedure and its drawbacks

The Swait and Louviere (1993) test, henceforth the SL-test, represents a likelihood ratio test comparing the preference structure across datasets. The data underlying each of the sixteen models is treated as a separate 'dataset'. Within sample preference dynamics are tested by applying the SL-test to two 'datasets' from the same sample at different moments along the choice sequence. Between sample dynamics are analysed in the SL-test by contrasting two 'datasets' at exactly the same moment in the choice sequence, but taken from a different sample. If two datasets have a similar preference structure, they can be merged, if not they need to be analysed separately.[5] The hard line between combining two datasets or treating them as independent introduces a trade-off between bias and efficiency. Merging the datasets may result in biased welfare estimates due to neglecting (subtle) differences in the underlying preference structure. Treating the two datasets as independent decreases the efficiency of parameter and welfare estimates.

The SL-test has its limitations when testing for within and between sample preference dynamics. Its binary approach to treating datasets as identical or independent implies that the test does not take into account that preferences at a particular stage of the choice sequence are more

[5] Alternative applications have contrasted data from revealed and stated preference studies (Adamowicz et al. 1994; Brownstone et al. 2000; Cameron et al. 2002) or compared welfare estimates across different populations in benefits transfer studies (Colombo et al. 2007; Johnston 2007; Lusk et al. 2003). Dynamics in the scale parameter over choice sequences have also been analyzed, which is commonly interpreted as a measure of choice accuracy (e.g. Brown et al. 2008; Kingsley and Brown 2010; Swait and Adamowicz 2001).

likely to be comparable to preferences revealed in choice tasks in its direct vicinity than to preferences revealed at the other end of the choice sequence. In fact, CA and DPH predict that preferences gradually evolve over the choice sequence before stabilizing at a specific level. In the next subsection a novel model is proposed, the local multinomial logit model (L-MNL), providing an intermediate solution between either combining different datasets or not and treating them as independent. The L-MNL model takes into account that preferences may gradually evolve over the choice sequence.

The local multinomial logit model

In the pooled multinomial logit (MNL) model marginal utility β is assumed to be constant across respondents and over the choice sequence (Holmes and Boyle, 2005). The interest of this paper is, however, in sample $s=1,2,...,S$ and choice task $t=1,2,\ldots,T$ specific preference parameters β_{st}. These can be obtained by estimating $S \cdot T$ independent models, which suffer from the same efficiency problems underlying the SL-test procedure. That is, the SL-test uses the independent models as inputs for the likelihood ratio test. The L-MNL model increases efficiency by estimating β_{st} whilst using information from all available data, i.e. from both samples and all eight choice tasks. Specifically, sixteen weighted MNL models are estimated: eight unique models within the LSB and HSB sample. These are labelled here as the locally estimated models.[6]

Each locally estimated model results in a vector of parameter estimates $\hat{\beta}_{st}$ for the respective local point (choice task t in sample s) based on the local log-likelihood function in

[6] Local likelihood estimation is discussed in Fan et al. (1995), and, for example, applied by Frölich (2006) and Fosgerau (2007).

Equation (1). Let y denote the vector of observed choices for all individuals $i=1,2,\ldots,I$ across all T choice tasks. X represents the matrix of associated explanatory variables in the utility function. The linear-additive utility specification results in the standard multinomial logit choice probability that individual i from sample q selects alternative j in choice task l: $P\left(y_{qil} = j \mid X_{qil}, \beta_{st}\right) = \frac{\exp\left(X_{qijl}\beta_{st}\right)}{\sum_{k=1}^{J}\exp\left(X_{qikl}\beta_{st}\right)}$. Finally, let I_q in the superscript in the second summation term on the right hand side of Equation (1) denote the number of respondents in sample q, being either the HSB or the LSB sample.[7]

$$LL_{st}(y|\beta_{st}, X) = \sum_{q=1}^{S}\sum_{i=1}^{I_q}\sum_{l=1}^{T} K_{ql} \cdot ln\left(P\left(y_{qil} = j \mid X_{qil}, \beta_{st}\right)\right) \tag{1}$$

The key element of the local likelihood function is formed by K_{ql} assigning a weight to each observation in the dataset. The subscript ql denotes that each choice task within each sample receives a unique weight. The weight is defined by the distance, i.e. degree of similarity, between each observation and the local point. Observations that are considered more similar to the local point, by being in the same sample ($q=s$) or by being positioned at the same moment in the choice sequence ($l=t$), receive a higher weight and therefore have more influence on the weighted log-likelihood function.

Formally, K_{ql} is determined by a kernel density function $g(\cdot)$, which requires as inputs: (i) a vector (or matrix) Z_{st} characterizing the local point; (ii) the value of Z at a specific observation

[7] The subscripts q and l are comparable to the subscripts s and t, but a change of notation is introduced because for each sample s and choice task t a unique model is estimated using all data from each sample q and each choice task l. The local point varies across models and thereby affects the weight of each observation in the likelihood function.

Z_{ql}; and (iii) a set of bandwidth parameters h, such that $K_{ql}=g(Z_{st}, Z_{ql}, h)$. Within- and between-sample preference dynamics are controlled for by means of a two-dimensional kernel density function, modelled as the product of two independent kernel density functions K_l^1 and K_q^2 in Equation (2).

$$K_{ql} = K_l^1 \cdot K_q^2 \tag{2}$$

$$K_l^1 = \begin{matrix} 1 \; if \; l = t \\ h_1^{|l-t|} if \; l \neq t \end{matrix} \tag{3}$$

$$K_q^2 = \begin{matrix} 1 \; if \; q = s \\ h_2 \; if \; q \neq s \end{matrix} \tag{4}$$

The first kernel density function in Equation (3) is associated with an ordered categorical variable. Specifically, K_l^1 describes a declining weight as choice task l moves further away from the local choice task t, for bandwidth parameter $h_1<1$. It relates to within-sample preference dynamics by assuming that choice tasks at the other end of the choice sequence are less likely to be based on the same utility function as choice tasks in the close proximity of t. K_q^2 is associated with an unordered categorical (dummy) variable defining to which sample (LSB or HSB) the local point s and an observation belongs. If q, the sample to which the observation belongs, is equivalent to the sample of the local point, it is assumed that the resulting choices are more likely to come from the same preference relation than for choices obtained from an alternative sample ($q \neq s$), for bandwidth parameter $h_2<1$. As such, K_q^2 relates to between sample preference dynamics. Racine et al. (2006) show that kernel density functions associated with ordered and unordered categorical variables need to have the possibility to be an indicator function; and that it must be possible to smooth out a categorical variable. The shape of the two kernel density functions fulfil these requirements when the bandwidth parameters are restricted to the interval [0,1].[8]

The bandwidth parameters smooth the locally estimated preference parameters. Non-zero bandwidth parameters are expected to result in an increase in efficiency relative to the set of sixteen independent models ($h_1 = h_2 = 0$), because it draws information from all observations in the dataset. If the bandwidth parameter is too large, then there is a risk of over-smoothing. Too much detail disappears and parameter estimates may become biased. If the bandwidth is too small, then there is a risk of under-smoothing, i.e. over-fitting due to random fluctuations in the data. A grid search is performed to identify the optimal set of bandwidth parameters h_1, h_2. Lower bandwidth parameters improve model fit, but introduce additional parameters in the model. In this paper, the corrected Akaike Information Criterion (*AICc*) is applied as a model selection criterion introducing a penalty for these additional parameters.[9]

The L-MNL model and the SL-test are comparable in the sense that both methods perform a preference structure test. The SL-test performs a likelihood ratio test to find out whether allowing for variation in preference parameters across stages (within and /or between samples) results in an improvement in model fit. The L-MNL does the same thing by optimizing the selected information criterion conditional on the (local) preference and bandwidth parameters. By smoothing the preference parameters, the L-MNL offers a more flexible and intermediate approach to the SL-test. Its bandwidth parameters are informative on the extent to which

[8] This implies that the kernel density can take the value $h=0$ for observations different than the local point. Other observations than the local point are not treated in the estimation of the L-MNL model. $h=1$ accounts for the fact that within or between sample preference dynamics may not be present. Specifically, a pooled dataset with the same β for each local point is obtained when setting ($h_1 = h_2 = 1$).

[9] The number of parameters in the model can be approximated by evaluating the trace of the hat-matrix (see Appendix B). Hurvich et al. (1998) provide a discussion on the use of alternative information criteria. As a rule of thumb, models are considered significantly different if the difference between model criteria is larger than $3/N$, where N is the total number of choices in the dataset (Charlton and Fortheringham 2009).

decisions at various stages of the choice sequence can be treated as similar. When preference dynamics are detected, both methods require statistical tests to find out whether the dynamics in the preference structure have implications for the welfare measures of interest. These tests require the comparison of parameters across local points, or independent models in case of the SL-test procedure. Since the scale parameter may vary across local models, scale-free marginal WTP estimates for specific attributes in the choice experiment are used as the basis for comparison in this paper.

IV. Empirical application

Flood risks in the Netherlands

Large parts of the Netherlands (26%), especially in the west, are situated below sea level and are threatened by an increase in coastal flood risks due to climate change (PBL 2010). Although most Dutch citizens know they live below sea level, they are generally not familiar, and have little to no experience, with making trade-offs regarding flood safety.[10] The central government and local water authorities have traditionally been responsible for providing and monitoring flood safety levels (Bouwer and Vellinga 2007). However, the Dutch government attempts to shift flood risk responsibilities from the public to the private sector as part of a broader cross-sectoral policy to make the country 'climate proof' (Kabat et al. 2005). Since there is currently a lack of incentives at the individual level to reduce exposure and vulnerability to flood risks, preferences are likely to be underdeveloped. Preference uncertainty may furthermore play a role as a result of the small

[10] Flood risk considerations in residential location choice are very limited (Brouwer and Schaafsma 2013).

probabilities associated with coastal flooding in the study area and the fact that most people never experienced a flood.[11]

A choice experiment is conducted in the densely populated western provinces of North-Holland and South-Holland, where major cities are located such as Amsterdam, The Hague and Rotterdam. The social and economic impacts of a coastal flood in this area are expected to be high. Some parts in the case study area are located almost six meters below sea level. The government aims to maintain a flood probability of once every 10,000 years in the area. Without additional investments in flood control, flood probabilities are expected to increase to once every 4,000 years by 2040 due to climate change (Maaskant et al. 2009). The interest of this paper is in the extent to which people are willing to trade-off an increase in their annual tax payments against a flood risk reduction by preventing the increase of the probability of a coastal flood and its associated socio-economic consequences.

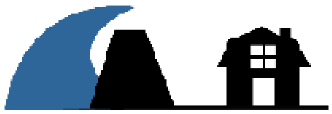
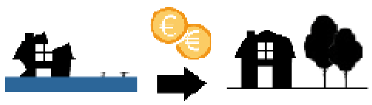
Survey administration

An online survey, conducted in March 2010, targeted a random selection of individual households in the two provinces, measuring their flood risk perception, flood preparedness and degree of risk aversion. Further details about the survey administration are provided in Dekker (2012). The key elements of the choice experiment embedded in the online survey are summarized here.

[11] The last catastrophic flood was in 1953 when more than 1,800 people died in the south-western part of the Netherlands.

TABLE 1

Attributes, attribute levels and definition of the status quo option

Attribute		*Possible attribute levels**			
Probability 	1 in 4,000 years	1 in 6,000 years (1.5x smaller)	1 in 8,000 years (2x smaller)	1 in 10,000 years (2.5x smaller)	
Compensation 	0%	50%	75%	100%	
Available evacuation time 	6 hours	9 hours	12 hours	18 hours	
Increase in annual tax 	€0	€40	€80	€120	€160

* The Status Quo alternative takes the most left (lowest) levels on all policy attributes

Two (unlabelled) alternative policies and a status quo (SQ) (opt-out) alternative are presented to the respondent. Each policy alternative is described by four attributes: (i) a reduction in flood probability; (ii) compensation of the material damage to each household after a coastal flood has occurred; (iii) available time to policy makers to prepare and conduct the evacuate the the study area before a flood occurs; and (iv) an increase in the annual tax to the water authority paid by all households, including the respondent's. Table 1 shows the design levels of each attribute and the definition of the SQ option. The relative size of the change in the probability compared to the SQ is also displayed on each choice card to increase public understanding of the associated magnitudes of the changes.

As described in Section II, a potential starting point bias is introduced in the first choice task. More specifically, respondents in the LSB were presented with the cost levels €40 and €80 for respectively the first and second alternative in the first choice card. These cost levels are €120 and €160 for the HSB sample. The policy alternatives depicted in the first choice task are identical for all other attribute levels in both samples. The remaining eight choice cards presented to the respondents in both samples come from the same experimental design.

V. Results

The sample consists of 477 respondents, respectively 247 in the HSB and 230 in the LSB sample. Together these respondents made 4,293 choices (477 times 9 choice tasks). Table 2 shows that the independent sampling strategy resulted in two sets of respondents comparable in terms of their main socio-economic characteristics. Statistical tests fail to reject the null-hypothesis of equivalence in the distribution and central tendency of these indicators across both samples. Given the comparability of the samples, a set of attributes-only multinomial logit models including a generic constant on the non-SQ alternatives, is presented to facilitate the illustration of the L-MNL model.[12]

TABLE 2

Testing for between sample equivalence in socio-economic sample characteristics

Variable	*Type*	*Description*	*Test*	*d.f.*	*Test-statistic*	*p-value*
Income	Categorical	10 (ordered) income categories	χ^2	9	8.52	0.48
Gender	Dummy	1= male ; 0 = female	χ^2	1	1.14	0.29
Age	Continuous	Respondent age (18-65)	Kolmogorov-Smirnov	-	0.08	0.50

The results section is structured in the following way. First, the benefits of the L-MNL model are illustrated by examining the issues of efficiency and bias associated with the set of

[12] If the samples are not comparable, it is not unlikely that variations in preferences due to uncontrolled heterogeneity are falsely attributed to within and between sample preference dynamics. Appendix C discusses results controlling for observed heterogeneity across respondents in the kernel density, and for unobserved heterogeneity in the utility function. The main conclusions are not affected by these more sophisticated model specifications. Finally, robust standard errors are used to correct for potential misspecification due to panel effects.

independent models and an overall pooled model neglecting preference dynamics. Second, the results of the L-MNL model are used to investigate the presence of a potential starting point bias and contrast the DPH and CA hypotheses. Third, the conclusions of the L-MNL model in terms of the DPH and CA hypotheses are contrasted with those of the Swait and Louviere test. The analysis is based on choice tasks 2-9, since the first choice task serves to induce the SPB.

The L-MNL model: Efficiency and Bias

Table 3 provides an overview of five alternative specifications of the L-MNL model. Model 1 sets both bandwidth parameters to zero implying that every local point is estimated independently, i.e. a unique MNL model for each choice task within each sample. Model 2 represents the opposite case in which all observations are pooled into a single MNL model by setting the bandwidth parameters to unity. The flexibility of Model 1 results in an improvement in log-likelihood compared to Model 2, but comes at the cost of introducing a large number of additional parameters.[13] The *AICc* reveals that Model 1 results in a worse outcome of the model selection criterion compared to Model 2. Table 4 provides insights into the efficiency problems associated with Model 1. The t-statistics associated with the marginal WTP estimates are rather low in general.[14] Particularly the probability and evacuation attributes show t-statistics below (or close to) 1.96 in more than 50% of the cases.

[13] Without specifying robust standard errors, there would be five parameters in each MNL model resulting in a total of 80 parameters for Model 1(16 independent models having 5 parameters each) and five for Model 2.

[14] WTP standard errors are based on the Delta-method (Hole, 2007).

TABLE 3

Overview of the AICc criterion for alternative L-MNL specifications

Model	*Description*	*Bandwidth* h_1 *(within)*	*Bandwidth* h_2*(between)*	*LL*	*Approx. # of pars*	*AICc*	*Average CoV*
(1)	Within + between sample variation	0.00	0.00	-3660.78	79.74	1.9616	0.549
(2)	MNL	1.00	1.00	-3720.78	10.63	1.9559	0.124
(3)	Optimal bandwidth parameter	0.43	0.20	-3677.24	24.01	1.9402	0.214
(4)	Optimal between sample variation	1.00	0.19	-3700.79	15.46	1.9480	0.151
(5)	Optimal within sample variation	0.46	1.00	-3703.98	16.04	1.9499	0.172

An improvement in t-statistics is observed when comparing the results of Model 1 with the results for Model 3 in Table 4. In Model 3, the *AICc* is optimized by controlling for within ($h_1=0.43$) and between ($h_2=0.20$) sample preferences dynamics. The *AICc* for Model 3 is significantly better compared to all other model specifications indicating that within and between preference dynamics are present in the database. The bandwidth parameters point out there is a degree of similarity between choices made at a comparable moment along the choice sequence and within the same database. By also drawing information from observations close to the local point, Model 3 is able to reduce the average coefficient-of-variation (CoV) for the WTP coefficient (see the average CoV in the final column of Table 3) by 61% compared to Model 1.[15] This is still less efficient than Model 2, which however runs the risk of biased parameter estimates. For completeness, Models 4 and 5 respectively control for between and within sample preference dynamics only. According to the average CoV, Models 4 and 5 are more efficient than Model 3 due to having less parameters in the model specification. However, they lack the flexibility to generate an improvement in the *AICc* compared to Model 3. Table 3 thereby supports the notion that within and between sample preference dynamics are present within the database.

[15] The CoV, also known as the signal to noise ratio, represents the ratio of the standard error to the mean of each WTP coefficient (four per model) and is averaged across all WTP estimates in the sixteen local models.

In order to test for a potential bias in WTP estimates between Models 2 and 3, a simple approach is applied. Assuming that WTP distributions are independently and normally distributed across the two models, the difference in WTP estimates between both models and the associated t-statistics are calculated.[16] Table 5 reveals both positive and negative deviations when comparing the marginal WTP estimates of Model 3 with the generic WTP estimates from Model 2. This is consistent with Model 2 being a smoothed version of Model 3. Support for a bias in Model 2 are found for the marginal rate of substitution between the alternative-specific-constant (ASC) and the cost attribute at the start of the choice sequence in the HSB sample. In accordance with predictions, respondents in the HSB sample are more willing to trade at the start of the choice sequence since the ASC is associated with the non-SQ policy alternatives. The bias decays rapidly and is no longer present after the third choice task. Similar biases at the attribute level are not observed. The results, however, confirm the concern of Holmes and Boyle (2005) that there is a risk of bias in pooling the data.

[16] The delta-method implies marginal WTP follows a normal distribution.

TABLE 4

Marginal WTP estimates based on L-MNL model 1

Sample	*HSB*								*LSB*							
	ASC		*Prob*		*Comp*		*Evac*		*ASC*		*Prob*		*Comp*		*Evac*	
Task	*Coeff*	*t-stat*	*Coeff*	*t-stat*	*Coeff*	*t-stat*	*Coeff*	*t-stat*	*Coeff*	*t-stat*	*Coeff*	*t-stat*	*Coeff*	*t-stat*	*Coeff*	*t-stat*
2	114.64	4.17	8.12	1.74	1.01	3.58	1.59	0.91	65.75	3.01	8.63	2.26	0.67	3.04	2.51	1.52
3	125.99	4.38	4.36	0.99	0.91	3.32	2.00	1.11	77.12	4.51	3.83	1.22	0.42	2.49	2.71	2.02
4	83.37	3.93	3.53	0.97	0.98	4.16	1.69	1.13	65.72	2.50	6.07	1.31	0.56	2.17	1.25	0.63
5	56.30	2.72	8.48	2.03	1.14	4.38	1.66	1.18	44.36	2.57	9.92	3.12	0.61	3.51	1.41	1.17
6	44.37	2.71	11.34	3.61	0.91	5.28	2.86	2.65	36.15	2.13	3.32	1.09	1.04	5.41	2.08	1.83
7	75.38	3.91	9.49	2.66	0.79	4.19	1.59	1.22	55.30	4.35	4.80	2.11	0.60	4.88	2.62	2.90
8	90.18	6.46	2.03	0.85	0.79	5.64	2.33	2.59	68.12	4.33	4.64	1.65	0.67	4.34	0.22	0.20
9	41.79	2.30	9.08	2.55	0.86	4.46	3.79	2.78	49.28	3.28	5.97	2.11	0.89	5.52	1.22	1.06

Marginal WTP estimates based on L-MNL model 3

Sample	*HSB*								*LSB*							
	ASC		*Prob*		*Comp*		*Evac*		*ASC*		*Prob*		*Comp*		*Evac*	
Task	*Coeff*	*t-stat*	*Coeff*	*t-stat*	*Coeff*	*t-stat*	*Coeff*	*t-stat*	*Coeff*	*t-stat*	*Coeff*	*t-stat*	*Coeff*	*t-stat*	*Coeff*	*t-stat*
2	98.68	7.39	6.78	2.97	0.92	5.99	1.86	1.93	72.20	6.00	6.86	3.38	0.66	5.28	2.27	2.59
3	93.86	8.46	5.67	3.06	0.89	6.70	1.97	2.31	72.13	7.11	5.59	3.34	0.61	5.80	2.20	3.04
4	77.62	7.77	5.93	3.51	0.92	7.85	1.89	2.60	63.75	6.25	6.17	3.78	0.67	6.28	1.79	2.68
5	63.61	6.96	7.54	4.76	0.94	8.68	1.98	3.22	53.55	6.19	7.03	4.87	0.73	7.92	1.76	3.20
6	58.63	6.97	8.20	6.00	0.90	9.05	2.28	4.17	50.38	6.36	5.81	4.49	0.81	9.70	1.98	3.97
7	66.21	7.70	7.25	5.40	0.83	8.59	2.18	4.13	56.07	7.60	5.40	4.45	0.73	10.04	2.01	4.18
8	70.94	8.81	5.36	4.21	0.82	8.92	2.29	4.33	60.89	7.56	5.07	3.83	0.74	9.21	1.42	2.68
9	58.69	6.20	6.71	4.08	0.84	7.83	2.75	4.07	55.19	6.28	5.68	3.71	0.80	8.64	1.51	2.36

ASC – Marginal rate of substitution between the cost attribute and the alternative specific constant

Probability – (€ per household per year for an extra 1,000 years in the denominator of the flood probability, from e.g. 1/4,000 → 1/5,000)

Compensation – (€ per household per year for an extra percentage point of compensation)

Evacuation – (€ per household per year for an extra hour of evacuation time)

TABLE 5

A test for bias between Models 2 and 3

	Sample	*HSB*								*LSB*							
		ASC		*Prob*		*Comp*		*Evac*		*ASC*		*Prob*		*Comp*		*Evac*	
	Task	*Coeff*	*t-stat*	*Coeff*	*t-stat*	*Coeff*	*t-stat*	*Coeff*	*t-stat*	*Coeff*	*t-stat*	*Coeff*	*t-stat*	*Coeff*	*t-stat*	*Coeff*	*t-stat*
Model 2	2-9	65.44	11.59	6.34	7.14	0.80	11.89	2.03	5.35	65.44	11.59	6.34	7.14	0.80	11.89	2.03	5.35
Difference	2	33.24	2.29	0.44	0.18	0.11	0.68	-0.16	-0.16	6.77	0.51	0.52	0.23	-0.14	-0.97	0.24	0.25
Model 3-2	3	28.43	2.28	-0.67	-0.33	0.09	0.58	-0.06	-0.06	6.70	0.58	-0.76	-0.40	-0.19	-1.52	0.17	0.21
	4	12.19	1.06	-0.41	-0.22	0.12	0.89	-0.14	-0.17	-1.69	-0.14	-0.18	-0.09	-0.13	-1.03	-0.23	-0.30
	5	-1.82	-0.17	1.19	0.66	0.14	1.10	-0.04	-0.06	-11.88	-1.15	0.68	0.40	-0.07	-0.61	-0.27	-0.40
	6	-6.80	-0.67	1.86	1.14	0.09	0.79	0.25	0.38	-15.06	-1.55	-0.54	-0.34	0.01	0.06	-0.04	-0.07
	7	0.77	0.07	0.91	0.56	0.03	0.24	0.16	0.24	-9.37	-1.01	-0.95	-0.63	-0.07	-0.74	-0.02	-0.03
	8	5.50	0.56	-0.98	-0.63	0.01	0.12	0.27	0.41	-4.54	-0.46	-1.28	-0.80	-0.06	-0.62	-0.60	-0.93
	9	-6.75	-0.61	0.36	0.19	0.04	0.29	0.72	0.93	-10.24	-0.98	-0.66	-0.37	0.00	0.00	-0.52	-0.70

TABLE 6

Testing if WTP in the LSB sample is higher than in the HSB sample using L-MNL model 3(p-values reported)

		P(WTP_LSB>WTP_HSB)		
Task	*ASC*	*Prob*	*Comp*	*Evac*
2	0.07*	0.51	0.10*	0.62
3	0.07*	0.49	0.05**	0.58
4	0.17	0.54	0.06*	0.46
5	0.21	0.41	0.07*	0.39
6	0.24	0.10	0.25	0.34
7	0.19	0.15	0.20	0.40
8	0.19	0.44	0.26	0.12
9	0.39	0.32	0.40	0.09*

*[**](***) indicates significance at the 10[5](1)% level

In summary, the L-MNL model offers a substantial improvement in efficiency compared to Model 1 whilst making dynamics in welfare estimates insightful. Indeed, the *AICc* supports the presence of within and between sample preference dynamics. At first sight these dynamics appear to be limited given that only limited evidence of a bias in welfare estimates between Models 2 and 3 is found. The next subsection provides a more detailed discussion regarding the observed within and between sample preference dynamics in the best-fit L-MNL model (Model 3) and the associated implications for contrasting the DPH and CA hypotheses.

The L-MNL model: SPB and DPH vs. CA

The different price vectors (for exactly the same policies) presented to the HSB and LSB sample in the first choice task result in the following choice patterns. Respondents in the first choice task tend to select the cheaper alternative. The share of SQ responses in the first choice task in the HSB sample (21%) is higher relative to the LSB sample (13%). The χ^2-test rejects the null hypothesis of an identical distribution of choice shares in the first choice task across the two subsamples at the 10% level (χ^2=5.52, p-value=0.06). As expected, the share of respondents selecting the SQ increases after the first choice task in the LSB sample and decreases in the HSB sample. This can be attributed to respondents being presented with respectively higher and lower prices for comparable flood risk reducing policies. Averaged over choice tasks 2-9, respondents in the LSB sample select the SQ option in 26% of the cases, while this share is 18% in the HSB sample. As such, the choice shares indicate the presence of a SPB.

Tables 4 and 6 point out that respondents in the HSB sample indeed reveal a lower tendency to select the SQ option compared to the LSB sample. This difference is significant in choice tasks two and three at the 10% level (Table 6). Moreover, marginal WTP for the compensation attribute is significantly higher in the HSB sample relative to the LSB sample until the sixth choice task. The L-MNL model thereby provides support, albeit limited, for the presence of a starting point bias due to anchoring on the price attribute in the first choice task.[17] The persistence of the SPB is further evaluated based on Table 6. After five choice tasks all welfare measures converge between samples. Only marginal WTP for the evacuation attribute is higher in the HSB sample in the final choice task, but only at the 10% significance level. These results are more in line with the DPH than the CA hypothesis. The impact of the initial choice task, and thus the starting point bias, wears off quickly. Stability of the preference relationship can, however, be questioned by the final choice task, requiring a closer look into the within sample preference dynamics as reported in Table 7.

[17] A status quo effect may be induced by what Loomes et al. (2009) label as taste uncertainty, where uncertain respondents exhibit trade-off resistance. Balcombe and Fraser (2011) also find that uncertain respondents have a higher propensity to defer from trading. Encountering lower prices in subsequent choice tasks as in the HSB sample may alleviate such a trade-off resistance possibly combined with learning effects.

TABLE 7

Summary of significant within sample preference dynamics based on L-MNL Model 3

Compare	*to*	*HSB*		*LSB*	
Task	*Task*	*ASC*	*PROB*	*ASC*	*COMP*
2	3				
2	4				
2	5	+**			
2	6	+***		+*	
2	7	+**			
2	8	+**			
2	9	+***			
3	4				
3	5	+**		+*	
3	6	+***		+**	-*
3	7	+**			
3	8	+**			
3	9	+***			-*
4	5				
4	6	+*			
4	7				
4	8				
4	9	+*			
5	6				
5	7				
5	8				
5	9				
6	7				
6	8		+*		
6	9				
7	8				
7	9				
8	9				

Note: Results for the other attributes are not reported, because no significant within sample preference dynamics were found.

*[**](***) indicates significance at the 10[5](1)% level

Within sample preference dynamics are more pronounced than between sample preference dynamics. Choice tasks two and three in the HSB sample reveal a higher ratio of the

ASC over the cost coefficient relative to choice tasks five to nine consistent with a lower tendency to select the SQ option at the start of the survey.[18] Significance levels are at least 5% in these cases. In choice task four, the SQ effect is still present compared to choice tasks six and nine, but only at the 10% significance level. The L-MNL model thereby supports a gradual decay of the SQ effect in the HSB sample indicating that the impact of the initial value cue reduces as respondents progress through the choice sequence. The SQ effect is one of the key findings in this paper. Regarding the other policy attributes, only a significant difference at the 10% level is found for the probability attribute when comparing choice tasks six and eight in the HSB sample. Within the LSB sample no such strong SQ effect is observed. Only in three instances a higher ASC to cost ratio is found when comparing the start of the choice sequence with later choice tasks.

Samuelson and Zeckhauser (1988) introduced the idea of a SQ bias, for which many researchers currently control in their analysis by specifying an error-components logit model (e.g. Meyerhoff and Liebe, 2009). We are not aware of other studies looking into the dynamics of this SQ bias over the choice sequence. Figure 1 depicts the development in the ASC to cost ratio over the choice sequence in both samples. Respondents in general seem to be more willing to make trade-offs across policy attributes, i.e. have a lower tendency to select the SQ, at the start of the survey. The value clues provided to each sample in the initial choice task clearly translate into an increased willingness to trade in the HSB sample in choice tasks two and three. Accordingly, the SQ effect is amplified by anchoring of respondent on high prices in the initial choice task. Until the sixth choice task, respondents in the HSB sample reduce their willingness to trade and are increasingly inclined to select the SQ option. A similar, but less pronounced pattern is observed

[18] An increase in the alternative specific constant and reduction in the cost coefficient increase the reported ratio and imply a higher probability to select one of the proposed policy alternatives.

for the LSB sample. By the end of the choice sequence, the ASC to cost ratio appears to recover and stabilises within both samples showing convergence.

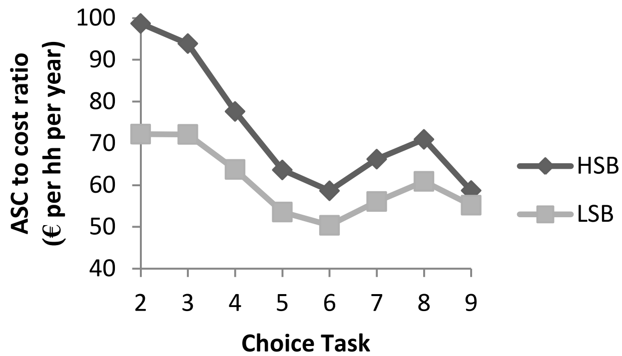


Figure 1: ASC to cost parameter ratio in both samples

In summary, the L-MNL model provides support for the DPH hypothesis. Within sample dynamics reveal that primarily choice tasks two and three are affected by the starting point bias. After choice task three, the observed SQ effect gradually wears off and significant differences in marginal WTP across policy attributes are only observed incidentally. In fact, marginal WTP estimates show signs of convergence between samples as predicted by the DPH. The robustness of the results was also tested. A set of sensitivity tests using respectively gender as an additional variable in the kernel density function, and a discrete random parameter on the cost coefficient is presented in Appendix C. These results do not affect our main conclusions.

The L-MNL model: L-MNL vs. SL-test

The previous two subsections pointed out that the L-MNL model offers improvements in terms of efficiency compared to using a set of independent models. Within and between sample preference dynamics were detected based on the new estimation method indicating that well-defined preferences are not available at the start of the choice sequence, making respondents vulnerable to arbitrary value cues. These effects gradually disappear as respondents repeatedly make similar decisions. This pattern confirms the DPH. In this section, we test if the same conclusions are obtained when using a set of independent models and the SL-test procedure.

Similar to the L-MNL model, the set of independent models also finds significant differences in the ASC to cost ratio between the two samples in choice tasks two and three (see Table D.1 in Appendix D). For the other policy attributes a more erratic pattern is observed consistent with the difference in parameter estimates between the top and bottom part of Table 4. For example, WTP for the compensation attribute is higher in the HSB sample in choice tasks three and five, whereas the L-MNL model found a more consistent effect across the first five choice tasks due to smoothing. Similarly, WTP for the probability attribute is higher in the HSB sample in choice tasks six and WTP for the evacuation attribute is significantly higher in tasks eight and nine. A consistent message regarding convergence in preferences between samples is thus not obtained based on the independent models. Within sample preference dynamics confirm the same erratic pattern precluding clear conclusions in contrasting the DPH and CA hypotheses. This comparison makes clear that the independent models are highly responsive to random fluctuations in preference patterns in particular choice tasks, also known as under-smoothing. The latter can be filtered out by increasing the number of observations per choice task or applying the smoothing procedure of the L-MNL model.

Rather than smoothing across all choice tasks, the SL-test treats two datasets as identical or independent. Table 8 presents the results of the SL-test. Columns two and three represent the log-likelihood of the choice task and sample specific models, of which the sum (column four), here across samples, is contrasted with the log-likelihood of a model imposing an identical preference structure across the two datasets. The fifth column allows for scale differences across the two datasets conditional on an equivalent preference structure, whereas the sixth column also imposes equality of scale. The SL-test does not support a starting point bias. In fact, columns seven and eight highlight that the preference structure is equivalent in the HSB and LSB sample in all choice tasks except tasks three and seven.[19] Also limited within sample preference dynamics are identified by the SL-test (see Tables D.3 and D.4 in Appendix D). The LSB sample does not reveal any differences in preference structure over the choice sequence, while in the HSB sample only significant differences are found between respectively choice tasks two, three and choice tasks six and nine. Apart from these erratic deviations, the SL-test neither confirms the DPH nor the CA hypothesis and is supportive of the standard micro-economic assumptions of stable preferences.

TABLE 8

Results for the between sample SL-test

Task	*LL HSB*	*LL LSB*	*LL SUM*	*LL Pooled rescaled*	*LL pooled*	*LR-test1*	*p-value*	*LR-test2*	*p-value*	*scale ln(HSB / LSB)*
2	-234.33	-232.03	-466.36	-468.94	-470.09	5.17	0.27	2.29	0.13	0.33
3	-238.44	-231.20	-469.64	-475.27	-476.07	11.27	0.02**	-	-	-
4	-238.93	-242.19	-481.11	-482.71	-486.01	3.20	0.53	6.59	0.01**	0.66
5	-236.19	-225.98	-462.17	-466.00	-466.28	7.66	0.10	0.55	0.46	0.16
6	-222.35	-219.96	-442.31	-445.93	-446.89	7.23	0.12	1.93	0.17	0.26
7	-232.57	-209.10	-441.67	-446.91	-447.26	10.48	0.03**	-	-	-
8	-217.26	-221.97	-439.23	-442.88	-444.74	7.30	0.12	3.71	0.05*	0.36
9	-241.60	-216.69	-458.29	-459.98	-460.21	3.38	0.50	0.47	0.49	-0.14

LR-test1 – Test for differences in the preference parameters, 4 degrees of freedom
LR-test2 – Test for differences in the scale parameter, 1 degree of freedom
*(**)[***] indicates significance at the 10(5)[1]% level

[19] In those cases where the hypothesis of equal preference parameters is rejected, it is meaningless to estimate the second LR-test statistic, and this test statistic is therefore also not presented in Table 8.

The independently estimated models used as inputs to the SL-test are inefficient and the procedure clearly supports smoothing the parameter estimates by fully combining most samples, which prevents researchers from testing for (subtle) preference dynamics in the welfare measures of interest. As such, the binary approach forces the SL-test to over-smooth the parameter estimates. The only form of flexibility present in the SL-test is controlling for scale differences across datasets. Significant scale differences between the two samples are found in choice tasks four and eight, where the HSB sample is found to have a higher scale parameter. Table 6 confirmed that choice task eight does not display a difference in welfare estimates between the two samples, but in choice task four marginal WTP for the compensation attribute is significantly lower in the LSB sample. In the SL-test, the inefficiency for choice task four is thus transferred into the scale parameter. Given the limitations of the SL-test at the current sample size, it is likely that the SL-test over-smoothes the within (and between) sample preference dynamics, an effect which may have consequently been picked up by the scale parameter. Indeed, significant within sample differences in the scale parameter are identified more often (see Tables D.3 and D.4).

Overall, the L-MNL proves to be a proper intermediate method falling in between the estimation of independent model and the SL-test procedure. Specifically, by controlling the degree of smoothing the method is able to detect preference dynamics while filtering out random fluctuations in preference parameters across choice tasks. This provides a more consistent picture of preference dynamics within and between samples without neglecting subtle preference dynamics. As a result, the method provides different conclusions with respect to contrasting the DPH and CA hypotheses. Where the SL-test procedure detects more or less stable preferences, the L-MNL model finds more support for the DPH.

VI. Conclusions

The existence of a set of well-defined preferences in many environmental economic valuation studies has been questioned due to unfamiliarity and inexperience of respondents with the policy attributes. Plott's (1996) discovered preference hypothesis and Ariely et al.'s (2003) coherent arbitrariness provide contradicting hypotheses on the extent to which respondents cope with this preference uncertainty and how preferences evolve over a sequence of choices. In this paper, the presence of between and within sample preference dynamics is examined in the face of an arbitrarily induced starting point bias in a hypothetical choice experiment. To this end, a uniquely designed stated choice survey on flood risk valuation is applied in combination with a new econometric model, which is considered to be better suited to test for gradual changes in preferences over a choice sequence. The developed model is contrasted with the Swait and Louviere (1993) test procedure, the most common approach to test for preference dynamics. In this paper we argue that the latter test procedure is less suitable to test for subtle dynamics in welfare estimates, in particular when sample sizes are considered small, which is usually the case for choice experiments.

The results of this paper are in line with findings by Ladenburg and Olsen (2008) and support the discovered preference hypothesis. Limited support for the existence of a (persistent) starting point bias in the choice experiment is found. The sample provided with a higher bid vector at the start of the choice sequence has a lower tendency to select the status quo option in subsequent choice tasks and thereby reveals a lower cost sensitivity. The impact of the initial choice task gradually disappears after the third choice task, resulting in a set of stable marginal WTP estimates in both samples, also in the sample given a lower bid at the start. More specifically, after the fifth choice task welfare estimates are no longer statistically different, at the

5% significance level, across the two samples in our novel local multinomial logit (L-MNL) model. On the contrary, the SL-test procedure tends to smooth out these preference dynamics at the start of the choice sequence.

Our main finding, the presence of a status quo effect and its associated dynamics, can be related to the lack of incentive compatibility in stated choice experiments (Carson and Groves, 2007). Respondents in the LSB sample were presented with similar alternatives at low(er) prices in the initial choice task and they may choose strategically in subsequent choice tasks by rejecting more expensive alternatives by selecting the status quo. Respondents in the HSB sample only observed high prices and are therefore more willing to trade initially. Gradually, more and more respondents in the HSB sample are also presented with cheaper alternatives and revert to the same strategic behaviour as in the LSB sample by becoming less willing to trade. The lack of incentive compatibility thus explains the general and sample specific decline in willingness-to-trade over the choice sequence. The former can, however, also be related to declining attention to the survey by the respondent. A closer examination into the dynamics of the status quo effects form an interesting topic for future research.

Four implications follow from this study. First, researchers should be aware of potential dynamics in welfare estimates over the choice sequence and not only focus on inherent differences in preferences across respondents (e.g. Hess and Rose 2009). Absence of stable welfare estimates in choice experiments complicates welfare analysis as it becomes unclear how many and which choice tasks should be used to this end. The only benchmark we can think of here is to compare these stated preferences to choices in a revealed preference setting to test if preference dynamics are an experimental artefact or explain real choice processes. Second, the

Swait and Louviere (1993) test procedure has the tendency to over-smooth the data and may thereby neglect possible dynamics in preferences when the underlying models are inefficient. The local MNL model is in our view better equipped for the purpose of testing for preference dynamics, because it offers improvements in flexibility and efficiency when estimating choice task specific preference parameters. Large reductions in standard errors are observed without the need to bundle observations from various choice cards. As such, the model is able to control for gradual changes in preferences and prevents against over-identification due to random variations in the data by smoothing parameter estimates. Applications of the local MNL model are furthermore not restricted to variations in preferences over time, but also across respondents. Third, additional effort needs to be placed in the development of experimental set-ups in which sample sizes and the experimental design enable researchers to estimate choice task specific choice models. Sample sizes used in this paper are comparable to those used in other studies, for example by Braga and Starmer (2005) or Ladenburg and Olsen (2008), who also use around 250-300 respondents per sample. Closely related, - and despite our careful study set-up -, individual respondents could have caused the observed dynamics in preferences, since at each moment in the sequence each choice card was answered by ten respondents on average. Therefore two sensitivity tests are conducted to additionally control for heterogeneity in preferences across respondents. The first test follows Ladenburg and Olsen (2008) and identifies whether the observed starting point bias is gender specific. The starting point bias is more apparent for male respondents, but still wears off after a couple of choice tasks. For the second test, a mixed logit model is estimated at the optimal bandwidth parameters of our local MNL model. A discrete distribution is imposed on the cost parameter allowing for unobserved heterogeneity in the cost parameter across respondents. The results are reported in Appendix C and confirm the main conclusions of this paper. Finally, the sensitivity of marginal WTP estimates to arbitrary initial value clues asks for careful testing of choice experiments and specification of the initial choice task. Looking beyond the scope of the current paper, an alternative approach would be to present respondents with an overview of all possible attribute levels before introducing a specific instructional choice task. In that case, starting point biases (or anchoring effects) may be circumvented by not presenting a single set of arbitrary value cues to the respondent (e.g. Bateman et al. 2004). This is actually not uncommon in practice. Since respondents are presented with all attribute levels, their tendency to select, for example, the status quo is more likely to be driven by the choice task at hand. However, the appropriateness of the levels included in the choice experiment needs to be defined in pre-testing stages while taking into account the preference uncertainty of respondents in those stages.

Appendix A – Experimental design: rotating and response frequencies

The experimental design consists of three blocks of 8 choice cards each. The total set of 24 choice cards is generated by Ngene (NGENE 2010) using a d-efficient design based on a random parameters error components logit model using 100 Halton draws (Rose and Bliemer 2009; Train 2009). The three non-cost attributes are assigned a normal distribution and the error component is used to control for a possible Status Quo effect (Scarpa et al. 2005). Non-zero priors applied in the design generation stage are based on pre-test results. Additional restrictions are imposed on the design to ensure that (i) the instructional choice tasks included in the LSB and HSB samples are not repeated in the subsequent choice sequence; (ii) no dominant alternatives are included in the choice sets; and (iii) the status quo alternative is not repeated as a policy alternative. Both the LSB and HSB sample are presented with the same set of choice cards after the first (instructional) choice task.

The three blocks of 8 choice cards are used to form 24 versions of the design. In order to optimize the estimation of a choice model at each moment in the choice sequence, the starting card rotates across versions. That is, Version 1 presents respondents with choice cards 1-8 in ascending order. Version 2 starts with choice cards 2-8 and ends with choice card 1. This rotation procedure yields 24 versions in total. Finally, the order of appearance of the first and second policy alternatives is altered to prevent effects from reading from left to right. Accordingly, the number of versions doubles to 48 and respondents are randomly assigned one version.

Table A.1 shows the number of times each block of the design is applied in both samples and the minimum number of times each block is fully answered, by different respondents, at each moment in the choice task. As such, the rotating procedure results that, on average, each choice card in the design is answered ten times at each moment in the design by respondents from a particular sample. A more detailed over view of response frequencies is provided in Tables A.2 and A.3. By evaluating the full design in each choice task, the model can be estimated more accurately at each moment in the choice sequence and results are not influenced by design elements. Our study differs in this respect from Ladenburg and Olsen (2008) who did not apply a similar rotating procedure and let all respondents answer the same choice task at the same moment during the choice sequence.

TABLE A.1

Number of times each block of the design is applied in the HSB and LSB samples

	# of times applied in each sample		*Minimum # of times fully applied in each choice task*	
	HSB	*LSB*	*HSB*	*LSB*
Block 1	86	77	8	5
Block 2	78	70	5	5
Block 3	83	83	7	8

TABLE A.2

Number of times each choice card in the design is applied during the choice sequence in HSB sample

HSB		*Task 2*	*Task 3*	*Task 4*	*Task 5*	*Task 6*	*Task 7*	*Task 8*	*Task 9*
Block 1	Card 1	8	12	12	12	9	11	11	11
	Card 2	11	8	12	12	12	9	11	11
	Card 3	11	11	8	12	12	12	9	11
	Card 4	11	11	11	8	12	12	12	9
	Card 5	9	11	11	11	8	12	12	12
	Card 6	12	9	11	11	11	8	12	12
	Card 7	12	12	9	11	11	11	8	12
	Card 8	12	12	12	9	11	11	11	8
Block 2	Card 9	11	11	9	10	5	12	12	8
	Card 10	8	11	11	9	10	5	12	12
	Card 11	12	8	11	11	9	10	5	12
	Card 12	12	12	8	11	11	9	10	5
	Card 13	5	12	12	8	11	11	9	10
	Card 14	10	5	12	12	8	11	11	9
	Card 15	9	10	5	12	12	8	11	11
	Card 16	11	9	10	5	12	12	8	11
Block 3	Card 17	13	9	9	11	12	11	11	7
	Card 18	7	13	9	9	11	12	11	11
	Card 19	11	7	13	9	9	11	12	11
	Card 20	11	11	7	13	9	9	11	12
	Card 21	12	11	11	7	13	9	9	11
	Card 22	11	12	11	11	7	13	9	9
	Card 23	9	11	12	11	11	7	13	9
	Card 24	9	9	11	12	11	11	7	13

TABLE A.3

Number of times each choice card in the design is applied during the choice sequence in LSB sample

LSB		*Task 2*	*Task 3*	*Task 4*	*Task 5*	*Task 6*	*Task 7*	*Task 8*	*Task 9*
Block 1	Card 1	11	10	10	11	9	11	10	5
	Card 2	5	11	10	10	11	9	11	10
	Card 3	10	5	11	10	10	11	9	11
	Card 4	11	10	5	11	10	10	11	9
	Card 5	9	11	10	5	11	10	10	11
	Card 6	11	9	11	10	5	11	10	10
	Card 7	10	11	9	11	10	5	11	10
	Card 8	10	10	11	9	11	10	5	11
Block 2	Card 9	12	11	9	7	5	9	7	10
	Card 10	10	12	11	9	7	5	9	7
	Card 11	7	10	12	11	9	7	5	9
	Card 12	9	7	10	12	11	9	7	5
	Card 13	5	9	7	10	12	11	9	7
	Card 14	7	5	9	7	10	12	11	9
	Card 15	9	7	5	9	7	10	12	11
	Card 16	11	9	7	5	9	7	10	12
Block 3	Card 17	8	9	12	11	11	9	10	13
	Card 18	13	8	9	12	11	11	9	10
	Card 19	10	13	8	9	12	11	11	9
	Card 20	9	10	13	8	9	12	11	11
	Card 21	11	9	10	13	8	9	12	11
	Card 22	11	11	9	10	13	8	9	12
	Card 23	12	11	11	9	10	13	8	9
	Card 24	9	12	11	11	9	10	13	8

Appendix B – Optimal bandwidth parameters

Fosgerau (2007) and Fröhlich (2006) argue that the bandwidth parameter generally has a larger impact on model results than the shape of the continuous kernel density itself. They also note that there is not a single bandwidth selection method considered to be the best. A practical approach is to select the smallest possible bandwidth for which all local models converge. This approach seems to work well for large datasets. However, it is unknown in advance if this will result in under-smoothing. Additional criteria are needed in order to have the possibility to test the model against the standard MNL model.

Hurvich et al. (1998) propose a statistic based on the trade-off between model fit and the number of parameters in the model, which can be used to determine the optimal bandwidth and select the appropriate model. The number of parameters in the model can be approximated by evaluating the trace of the hat-matrix H (see below). If the bandwidth h of a categorical variable is low, the fit of the model will be better, but more parameters are needed, so the trace of the hat matrix $tr(H)$ will be higher. Model evaluation criteria like the *Akaike information criterion* (AIC) and *Bayesian information criterion* (BIC) can be used for selecting the optimal bandwidth. Hurvich et al. (1998) note that the AIC can lead to under-smoothing, while the BIC tends to support a high degree of smoothing. In this paper, the corrected AIC (AICc) is applied as model selection criterion $AICc = \frac{-2LL\left(\hat{\beta}\right)}{I \cdot T} + \frac{2 \cdot tr\left(\hat{H}\right)+1}{I \cdot T - tr\left(\hat{H}\right)-2}$, introducing an additional penalty for additional parameters in the model compared to the AIC. As a rule of thumb, models are considered significantly different if the difference between model criteria is larger than $3/(I \cdot T)$ (Charlton and Fortheringham 2009).

As discussed in Koster and Koster (2013), the L-MNL method has its drawbacks if panel data are used. If one does not correct for the panel nature of the data, the local standard errors will be underestimated. Therefore, the trace of the hat-matrix becomes too low, which will result in an optimal bandwidth that is too low and therefore under-smoothing of the model. This paper correct for this by estimating robust standard errors clustered over respondents (Freedman 2006).

Nagel and Hatzinger (1992) and Koster and Koster (2013) are followed in deriving the hat-matrix for each of the $I \cdot T$ locally estimated weighted MNL models. Let Ω_l represent the $k \cdot k$ (robust) covariance matrix of parameter estimates belonging to a specific locally estimated weighted MNL model l. Alternatively, Ω_l can be specified as the inverse hessian matrix $\Omega_l=(X^{*\prime} V_l X^*)^{-1}$, but using the covariance matrix reduces computation time. X^* is a transformation of the design matrix X, where each observation is multiplied by the square root of its own weight $\sqrt{K_{it}}$.[20] V_l represents the locally estimated covariance matrix of choice probabilities. Due to the IIA property of the (weighted) MNL model, V_l is a block diagonal matrix containing the observation specific covariance matrices of estimated choice probabilities V_{it}^l along the main diagonal:

$$V_{it}^l = \begin{pmatrix} \hat{P}_1\left(1-\hat{P}_1\right) & \dots & -\hat{P}_{J-1}\hat{P}_1 \\ \vdots & \ddots & \vdots \\ -\hat{P}_1\hat{P}_{J-1} & \cdots & \hat{P}_{J-1}\left(1-\hat{P}_{J-1}\right) \end{pmatrix}$$

$$V^l = \begin{pmatrix} V_{11}^l & & 0 \\ & \ddots & \\ 0 & & V_{nT}^l \end{pmatrix}$$

[20] More formally, X is a $I \cdot T \cdot (J\text{-}1)$ by k matrix describing the characteristics of each alternative adjusted for a reference alternative (in our case the status quo option). Additionally, it also includes additional explanatory variables in the model. Hence, each observation is described by (J-1) rows in X.

Nagel and Hatzinger (1992) define the hat-matrix for a standard MNL model by $H=V^{1/2}X(X'VX)^{-1}X'V^{1/2}{}'$. This specification is used to construct the hat-matrix for the locally estimated weighted MNL model l. Rewriting $X^{*}{}'V_lX^{*}= X^{*}{}'V_l^{1/2}{}'V_l^{1/2}X^{*}$ and noting the similarity between this and the specification by Nagel and Hatzinger (1992), the local Hat-matrix in the following way: $H_l=V_l^{1/2}X^{*}(X^{*}{}'V_lX^{*})^{-1}X^{*}{}'V_l^{1/2}{}'$. The specification can be further simplified by replacing the middle statement by the local covariance matrix. $H_l= V_l^{1/2}X^{*}\Omega_l X^{*}{}'\ V_l^{1/2}{}'$. Note that for each local point a local Hat-matrix needs to be derived.

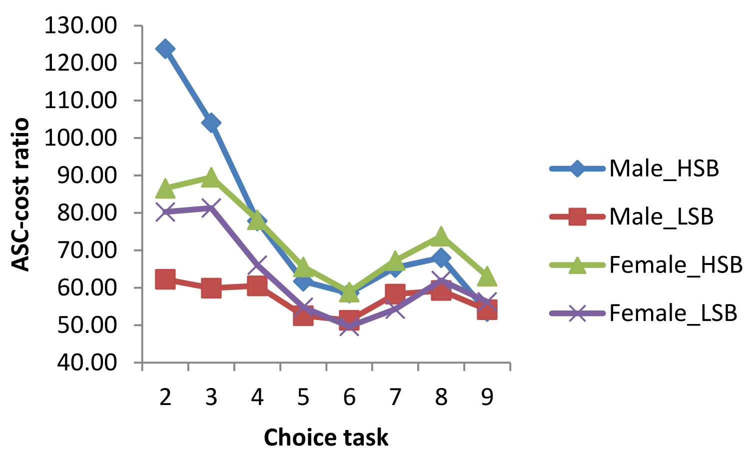
Using properties of linear algebra, the trace of the local Hat-matrix can be rewritten by $tr(H_l)=tr(X^{*}\Omega_l X^{*}{}'V_l)$, which saves substantial computation time. As mentioned in Section IV, the trace of the Hat-matrix approximates the number of parameters in the local model. In the eventual comparison of alternative bandwidth parameters, only the trace elements of the local hat matrix belonging to the local point are used and summed. More specifically, for the first choice card, which contains three alternatives in our case, the first two trace elements of the local hat matrix are stored. For local point two, the elements three and four from its own local hat-matrix. In order to reduce computation time, specific elements c on the trace of the local Hat-matrix can be obtained by calculating $X^{*}(c,:)\Omega_l X^{*}{}'V_l\ (:,c)$, picking the c-th row of X^{*} and the c-th column of V_l. The number of parameters related to a specific bandwidth parameter is approximated by summing the stored trace elements over all local models. Clearly, under uniform weights the hat-matrix reduces to the MNL hat-matrix in which the trace sums to the exact number of parameters in the model.

Appendix C – Sensitivity test controlling for preference heterogeneity across respondents

First a test for gender effects is conducted by additionally controlling for the binary variable 'Gender' in the kernel density function. The bandwidth parameter is optimized ($h_{gender}=0.26$) while keeping the bandwidth parameters for within and between sample preference dynamics constant (see Model 3 in Table 3). Figure C.1 shows that the starting point bias in the ASC-cost ratio is more apparent for male respondents. This result is in contrast to Ladenburg and Olsen (2008) who find that specifically females are significantly affected by the starting point bias. For the ASC-cost ratio, and also WTP for the compensation attribute, similar convergence patterns are observed across all four depicted subsamples. WTP estimates for the probability attribute are more ad hoc over the choice sequence in this model specification, but again WTP levels seem to converge between the samples over the choice sequence. Last but not least, the evacuation attribute still reveals a divergence of WTP estimates in the final choice task, but this effect turns out not to be gender specific. Hence, our support for the discovered preference hypothesis is not affected by this sensitivity test.[21]

FIGURE C.1

Gender and sample specific ASC-cost ratio over the choice sequence



[21] Detailed results for all policy attributes are available upon request from the corresponding author.

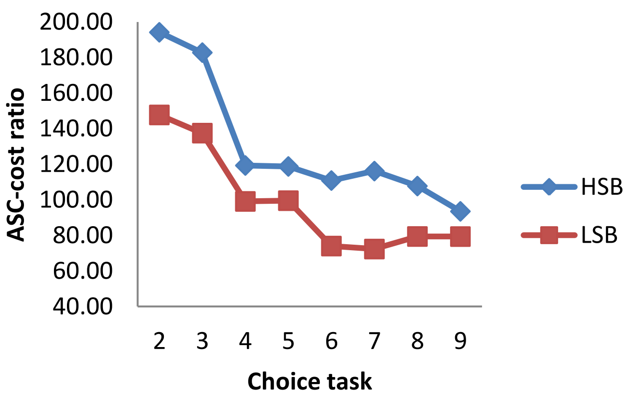
The second sensitivity test aims to control for unobserved heterogeneity across respondents using a cross-sectional latent class model with two classess. The choice probability of Equation (1) for local point *st* is then modified to:

$$P\left(y_{qil}=j \middle| p_{st},\beta_{st},X_{qil}\right)=p_{1st}\cdot\frac{exp(X_{qijl}\beta_{1st})}{\sum_{k=1}^{J_{qil}} exp(X_{qikl}\beta_{1st})}+(1-p_{1st})\cdot\frac{exp(X_{qijl}\beta_{2st})}{\sum_{k=1}^{J_{qil}} exp(X_{qikl}\beta_{2st})}$$

where $p_{1st} \in [0,1]$ is the probability to belong to the group with preference parameters $\beta_{1st}$and $1-p_{1st}$the probability to belong to the group with preference parametersβ_{2st}. For computational tractability a mixing distribution on the cost coefficient is estimated. Again, the model is estimated at the optimal bandwidth parameters as presented in Table 3. Figure C.2 shows again that the mean ASC to cost ratio decreases and that the effect of the starting point bias wears out, implying that the observed patterns of within and between sample preference dynamics are similar to the results obtained for the basic L-MNL model. Not surprisingly, WTP levels are affected as illustrated by the higher level of the ASC to cost ratio over the entire sequence in Figure C.2. Patterns for the other policy attributes and the ASC are available upon request and do not contradict the conclusions from the main text in terms of the patterns over the choice sequence.

FIGURE C.2

Mean ASC to cost ratio controlling for unobserved preference heterogeneity in the cost parameter



Appendix D – Results independent models and the SL-test

This appendix presents the results for within and between preference dynamics based on the set of independently estimated models (L-MNL model 1). It also provides more detail on the outcomes of the Swait and Louviere (1993) test procedure.

Table D.1 presents the comparison for between sample preference dynamics. Similar to L-MNL model 3, the set of independent models also find significant differences in the ASC to cost ratio between the two samples in choice tasks two and three. For the other policy attributes a more erratic pattern is observed. For example, WTP for the compensation attribute is higher in the HSB sample in choice tasks three and five, where the L-MNL found a more consistent effect across the first five choice tasks due to smoothing. A consistent message regarding convergence in preferences between samples can thus not be obtained based on the independent models.

TABLE D.1

Testing if WTP in the LSB sample is higher than in the HSB sample using L-MNL model 1(p-values reported)

	P(WTP_LSB>WTP_HSB)			
Task	*ASC*	*Prob*	*Comp*	*Evac*
2	0.08*	0.53	0.17	0.65
3	0.07*	0.46	0.06*	0.62
4	0.30	0.67	0.11	0.43
5	0.33	0.61	0.05**	0.45
6	0.36	0.03**	0.70	0.31
7	0.19	0.13	0.20	0.74
8	0.15	0.76	0.28	0.07*
9	0.62	0.25	0.54	0.07*

*[**](***) indicates significance at the 10[5](1)% level

Within sample preference dynamics confirm the same erratic pattern precluding clear conclusions in contrasting the DPH and CA hypotheses (see Table D.2). This comparison makes clear that the independent models are highly responsive to random fluctuations in preference patterns in particular choice tasks, also known as under-smoothing. The latter can be filtered out by increasing the number of observations per choice task or applying the smoothing procedure embedded in the L-MNL model.

TABLE D.2

Summary of significant within sample preference dynamics based on L-MNL Model 1

		HSB		*LSB*			
Task1	*Task2*	*ASC*	*PROB*	*ASC*	*PROB*	*COMP*	*EVAC*
2	3						
2	4						
2	5	+**					
2	6	+**					
2	7						
2	8						
2	9	+**					
3	4						
3	5	+**		+*	-*		
3	6	+***	-*	+**		-***	
3	7	+*					
3	8						+*
3	9	+***				-**	
4	5						
4	6	+*	-*			-*	
4	7						
4	8						
4	9	+*					
5	6				+*	-*	
5	7				+*		
5	8	-*	+*				
5	9						
6	7					+**	
6	8	-*	+***	-*		+*	
6	9						
7	8		+**				+**
7	9					-*	
8	9	+**	-*				

Note: Results for the other attributes are not reported, because no significant within sample preference dynamics were found.

*[**](***) indicates significance at the 10[5](1)% level

The Swait and Louviere test procedure consists of three stages. First, the researcher splits the sample into two alternative subsamples, in our case the samples HSB and LSB or specific choice tasks, and then estimates the unrestricted model with a set of unique preference parameters for each subsample. Second, a restricted model is estimated with a common set of preference parameters, but a varying (relative) scale parameter across the two subsamples.[22] A likelihood ratio test (LR-test 1) is applied to test equivalence of all preference parameters between two samples, with the degrees of freedom being equal to *k-1*, where *k* is the number of imposed parameter restrictions. One degree of freedom is lost by explicitly estimating the relative scale parameter. If the null-hypothesis of equivalent preferences is rejected, the samples cannot be combined and it is unknown whether the observed differences arise due to variation in preferences or also due to variations in scale. The third and final step is only conducted when the former null-hypothesis is not rejected. It tests whether scale is equivalent across both samples. A pooled model with common scale and preference parameters is estimated and its log-likelihood value is contrasted against the second stage model using a likelihood ratio test with one degree of freedom for restricting the relative scale parameter (LR-test 2). The null-hypothesis assumes scale is equivalent in both samples.

Tables D.3 and D.4 report the results regarding within sample preference dynamics based on the SL test procedure for respectively the HSB and LSB sample. Within the HSB sample only significant differences are found between respectively choice tasks two, three and choice tasks six and nine (LR-test 1 columns 8-9). Within the LSB sample no preference dynamics are detected based on LR-test 1. LR-test 2 (columns 10-12) points out that the scale parameter

[22] Due to the confounding between preference and scale parameters, variations in scale can only be retrieved after imposing equivalence of preference parameters. For identification normalization of a single scale parameter is required. We normalize the scale parameter to one.

increases towards the end of the choice sequence indicating learning effects. Differences in scale within the LSB sample are mainly found with respect to choice task four, but also choice tasks two and three report lower scale levels.Such a pattern is not observed for the HSB sample. Test for between sample preference dynamics are reported in Table 8 in the main text.

TABLE D.3

Summary of within sample preference dynamics in the HSB sample based on the SL-test procedure

Task 1	*Task 2*	*LL1*	*LL2*	*SUM*	*Pooled scale*	*Pooled*	*LR-test 1*	*p-value*	*LR-test 2*	*p-value*	*scale*
2	3	-234.33	-238.44	-472.77	-472.99	-473.06	0.45	0.98	0.14	0.71	0.93
2	4	-234.33	-238.93	-473.25	-474.75	-474.84	2.98	0.56	0.19	0.67	0.92
2	5	-234.33	-236.19	-470.52	-472.63	-472.64	4.21	0.38	0.04	0.85	0.96
2	6	-234.33	-222.35	-456.68	-460.70	-461.30	8.04	0.09*			
2	7	-234.33	-232.57	-466.89	-468.51	-468.51	3.22	0.52	0.00	0.95	1.01
2	8	-234.33	-217.26	-451.59	-454.43	-455.51	5.69	0.22	2.16	0.14	1.30
2	9	-234.33	-241.59	-475.92	-480.57	-480.76	9.29	0.05*			
3	4	-238.44	-238.93	-477.36	-478.70	-478.71	2.68	0.61	0.01	0.94	0.98
3	5	-238.44	-236.19	-474.63	-477.29	-477.30	5.31	0.26	0.03	0.86	1.03
3	6	-238.44	-222.35	-460.79	-465.17	-466.23	8.76	0.07*			
3	7	-238.44	-232.57	-471.00	-472.90	-472.98	3.79	0.43	0.16	0.69	1.08
3	8	-238.44	-217.26	-455.70	-457.59	-459.18	3.77	0.44	3.20	0.07*	1.38
3	9	-238.44	-241.59	-480.03	-484.58	-484.62	9.10	0.06*			
4	5	-238.93	-236.19	-475.12	-475.76	-475.79	1.28	0.86	0.06	0.80	1.05
4	6	-238.93	-222.35	-461.28	-463.22	-464.42	3.87	0.42	2.41	0.12	1.32
4	7	-238.93	-232.57	-471.49	-472.34	-472.45	1.69	0.79	0.22	0.64	1.09
4	8	-238.93	-217.26	-456.19	-456.70	-458.52	1.03	0.91	3.64	0.06*	1.40
4	9	-238.93	-241.59	-480.52	-482.56	-482.58	4.08	0.39	0.03	0.86	0.96
5	6	-236.19	-222.35	-458.55	-459.54	-460.37	1.99	0.74	1.66	0.20	1.26
5	7	-236.19	-232.57	-468.76	-469.53	-469.55	1.55	0.82	0.04	0.84	1.04
5	8	-236.19	-217.26	-453.45	-455.70	-457.10	4.50	0.34	2.79	0.09*	1.35
5	9	-236.19	-241.59	-477.79	-479.40	-479.48	3.22	0.52	0.17	0.68	0.92
6	7	-222.35	-232.57	-454.92	-455.75	-456.29	1.66	0.80	1.08	0.30	0.84
6	8	-222.35	-217.26	-439.62	-443.14	-443.21	7.06	0.13	0.13	0.72	1.06
6	9	-222.35	-241.59	-463.95	-464.37	-465.78	0.84	0.93	2.82	0.09*	0.74
7	8	-232.57	-217.26	-449.83	-451.54	-452.64	3.43	0.49	2.20	0.14	1.29
7	9	-232.57	-241.59	-474.16	-475.61	-475.82	2.90	0.57	0.41	0.52	0.88
8	9	-217.26	-241.59	-458.86	-461.36	-463.54	5.01	0.29	4.37	0.04**	0.68

LR-test1 – Test for differences in the preference parameters, 4 degrees of freedom

LR-test2 – Test for differences in the scale parameter, 1 degree of freedom

*(**)[***] indicates significance at the 10(5)[1]% level

TABLE D.4

Summary of within sample preference dynamics in the LSB sample based on the SL-test procedure

Task 1	*Task 2*	*LL1*	*LL2*	*SUM*	*Pooled scale*	*Pooled*	*LR-test 1*	*p-value*	*LR-test 2*	*p-value*	*scale*
2	3	-232.03	-231.20	-463.23	-464.55	-464.55	2.65	0.62	0.00	0.99	1.00
2	4	-232.03	-242.19	-474.22	-474.81	-475.94	1.19	0.88	2.26	0.13	0.66
2	5	-232.03	-225.98	-458.01	-459.59	-459.77	3.15	0.53	0.36	0.55	1.15
2	6	-232.03	-219.96	-451.99	-454.80	-455.58	5.63	0.23	1.56	0.21	1.33
2	7	-232.03	-209.10	-441.13	-442.97	-445.59	3.67	0.45	5.24	0.02**	1.60
2	8	-232.03	-221.97	-454.00	-455.84	-456.42	3.68	0.45	1.17	0.28	1.28
2	9	-232.03	-216.69	-448.72	-450.21	-451.33	2.97	0.56	2.25	0.13	1.38
3	4	-231.20	-242.19	-473.39	-473.90	-474.95	1.03	0.90	2.08	0.15	0.67
3	5	-231.20	-225.98	-457.18	-459.40	-459.67	4.45	0.35	0.53	0.47	1.18
3	6	-231.20	-219.96	-451.16	-454.89	-455.80	7.46	0.11	1.82	0.18	1.36
3	7	-231.20	-209.10	-440.30	-441.07	-443.76	1.54	0.82	5.38	0.02**	1.59
3	8	-231.20	-221.97	-453.17	-455.30	-455.98	4.27	0.37	1.35	0.25	1.30
3	9	-231.20	-216.69	-447.89	-450.86	-452.16	5.94	0.20	2.59	0.11	1.43
4	5	-242.19	-225.98	-468.17	-468.48	-470.75	0.63	0.96	4.54	0.03**	1.74
4	6	-242.19	-219.96	-462.15	-463.28	-467.06	2.27	0.69	7.56	0.01***	2.02
4	7	-242.19	-209.10	-451.29	-451.57	-458.56	0.56	0.97	13.99	0.00***	2.35
4	8	-242.19	-221.97	-464.16	-464.39	-467.55	0.46	0.98	6.33	0.01**	1.88
4	9	-242.19	-216.69	-458.88	-459.40	-463.86	1.03	0.90	8.93	0.00***	2.07
5	6	-225.98	-219.96	-445.94	-448.27	-448.45	4.67	0.32	0.36	0.55	1.14
5	7	-225.98	-209.10	-435.08	-436.52	-437.89	2.87	0.58	2.74	0.10	1.38
5	8	-225.98	-221.97	-447.95	-449.13	-449.24	2.37	0.67	0.22	0.64	1.10
5	9	-225.98	-216.69	-442.67	-443.91	-444.30	2.48	0.65	0.77	0.38	1.20
6	7	-219.96	-209.10	-429.06	-431.44	-432.01	4.75	0.31	1.14	0.28	1.23
6	8	-219.96	-221.97	-441.93	-443.61	-443.62	3.38	0.50	0.01	0.90	0.98
6	9	-219.96	-216.69	-436.65	-437.27	-437.30	1.25	0.87	0.06	0.80	1.05
7	8	-209.10	-221.97	-431.07	-432.98	-433.77	3.82	0.43	1.58	0.21	0.79
7	9	-209.10	-216.69	-425.79	-427.98	-428.34	4.38	0.36	0.71	0.40	0.86
8	9	-221.97	-216.69	-438.66	-439.33	-439.42	1.34	0.85	0.17	0.68	1.08

LR-test1 – Test for differences in the preference parameters, 4 degrees of freedom

LR-test2 – Test for differences in the scale parameter, 1 degree of freedom

*(**)[***] indicates significance at the 10(5)[1]% level

References

Adamowicz W., Louviere J. and Williams M. (1994). ‘Combining Revealed and Stated Preference Methods for Valuing Environmental Amenities’, *Journal of Environmental Economics and Management*, vol. 26, pp. 271-292.

Ariely D., Loewenstein G. and Prelec D. (2003). ‘Coherent Arbitrariness: Stable Demand Curves Without Stable Preferences’, *Quarterly Journal of Economics*, vol. 118, pp. 73-105.

Ariely D., Loewenstein G. and Prelec D. (2006). ‘Tom Sawyer and the construction of value’, *Journal of Economic Behavior & Organization*, vol. 60, pp. 1-10.

Balcombe K. and Fraser I. (2011). ‘A general treatment of don't know responses from choice experiments’, European Review of Agricultural Economics, vol. 38, pp. 171-191.

Bateman I. J., Burgess D., Hutchinson W. G. and Matthews D. I. (2008). ‘Learning design contingent valuation (LDCV): NOAA guidelines, preference learning and coherent arbitrariness’, *Journal of Environmental Economics and Management*, vol. 55, pp. 127-141.

Bateman I. J., Cole M., Cooper P., Georgiou S., Hadley D. and Poe G. L. (2004), ‘On visible choice sets and scope sensitivity’, *Journal of Environmental Economics and Management*, vol .47, pp. 71-93.

Bech M., Kjaer T. and Lauridsen J. (2011). ‘Does the number of choice sets matter? Results from a web survey applying a discrete choice experiment’, *Health Economics*, vol. 20, pp. 273-286.

Bouwer, L. M. & Vellinga, P. (2007). ‘On the Flood Risk in the Netherlands,’ in *Flood Risk Management in Europe*, S. Begum, M. J. F. Stive, & J. W. Hall (eds), Springer Netherlands, pp. 469-484.

Braga J. and Starmer C. (2005). ‘Preference Anomalies, Preference Elicitation and the Discovered Preference Hypothesis’, *Environmental and Resource Economics*, vol. 32, pp. 55-89.

Brouwer R., Dekker T., Rolfe J. and Windle J. (2010). ‘Choice Certainty and Consistency in Repeated Choice Experiments’, *Environmental and Resource Economics*, vol. 46, pp. 93-109.

Brouwer, R. And Schaafsma, M. (2013). Modelling risk adaptation and mitigation behaviour under different climate change scenarios. *Climatic Change*, 117: 11-29.

Brown T. C., Kingsley D., Peterson G. L., Flores N. E., Clarke A. and Birjulin A. (2008). ‘Reliability of individual valuations of public and private goods: Choice consistency, response time, and preference refinement’, *Journal of Public Economics*, vol. 92, pp. 1595-1606.

Brownstone D., Bunch D. S. and Train K. (2000). ‘Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles’, *Transportation Research Part B: Methodological*, vol. 34, pp. 315-338.

Cameron T. A., Poe G. L., Ethier R. G. and Schulze W. D. (2002). ‘Alternative Non-market Value-Elicitation Methods: Are the Underlying Preferences the Same?’, *Journal of Environmental Economics and Management*, vol. 44, pp. 391-425.

Carlsson F. and Martinsson P. (2008). ‘How Much is Too Much?’, *Environmental and Resource Economics*, vol. 40, pp. 165-176.

Carlsson, F., Morkbak, M. R., & Olsen, S. B. (2010). *The first time is the hardest: a test for ordering effects in choice experiments* Goeteborg University Working Paper in Economics no.470.

Carson R. and Groves T. (2007). ‘Incentive and informational properties of preference questions’, *Environmental and Resource Economics*, vol. 37, pp. 181-210.

Charlton, M. & Fortheringham, S. (2009). Geographically weighted regression: a white paper. http://ncg.nuim.ie/ncg/GWR/GWR_WhitePaper.pdf

Colombo S., Calatrava-Requena J. and Hanley N. (2007). ‘Testing Choice Experiment for Benefit Transfer with Preference Heterogeneity’, *American Journal of Agricultural Economics*, vol. 89, pp. 135-151.

Dekker, T. (2012). ‘Preference uncertainty in stated choice experiments’, *PhD Thesis,*VU University Amsterdam http://dare2.ubvu.vu.nl/handle/1871/38275

Fan J., Heckman N. E. and Wand M. P. (1995). ‘Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions’, *Journal of the American Statistical Association,* vol. 90, pp. 141-150.

Fischhoff B., Welch N. and Frederick S. (1999). ‘Construal Processes in Preference Assessment’, *Journal of Risk and Uncertainty,* vol. 19, pp. 139-164.

Fosgerau M. (2007). ‘Using nonparametrics to specify a model to measure the value of travel time’, *Transportation Research Part A: Policy and Practice,* vol. 41, pp. 842-856.

Freedman D. A. (2006). ‘On The So-Called Huber Sandwich Estimator and Robust Standard Errors’, *The American Statistician*, vol. 60, pp. 299-302.

Frohlich M. (2006). ‘Non-parametric regression for binary dependent variables’, *Econometrics Journal*, vol. 9, pp. 511-540.

Groeneveld R. (2010). ‘Framing and training to induce preference learning in choice experiments’, *Marine Resource Economics*, vol. 25, pp. 233-245.

Hanley N., Adamowicz W. and Wright R. E. (2005). ‘Price vector effects in choice experiments: an empirical test’, *Resource and Energy Economics*, vol. 27, pp. 227-234.

Hess S. and Rose J. M. (2009). ‘Allowing for intra-respondent variations in coefficients estimated on repeated choice data’, *Transportation Research Part B: Methodological,* vol. 43, pp. 708-719.

Hole A.R. (2007). ‘A comparison of approaches to estimating confidence intervals for willingness to pay measures’, *Health Economics*, vol. 16, pp.827-840.

Holmes T. P. and Boyle K. J. (2005). ‘Dynamic Learning and Context-Dependence in Sequential, Attribute-Based, Stated-Preference Valuation Questions’, *Land Economics,* vol. 81, pp. 114-126.

Hurvich C. M., Simonoff J. S. and Tsai C. L. (1998). ‘Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* vol. 60, pp. 271-293.

Johnston R. (2007). ‘Choice experiments, site similarity and benefits transfer’, *Environmental and Resource Economics*, vol. 38, pp. 331-351.

Kabat P., van Vierssen W., Veraart J., Vellinga P. and Aerts J. (2005). ‘Climate proofing the Netherlands’, *Nature*, vol. 438, pp. 283-284.

Kingsley D. C. and Brown T. C. (2010). ‘Preference Uncertainty, Preference Learning, and Paired Comparison Experiments’, *Land Economics*, vol. 86, pp. 530-544.

Koster, P. & Koster, H. (2013). ‘Commuters’ preferences for fast and reliable travel’, Tinbergen Discussion Paper.

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*, vol. 56, pp. 275-285.

Loomes G., Orr S. and Sugden R. (2009). ‘Taste uncertainty and status quo effects in consumer choice’, *Journal of Risk and Uncertainty*, vol. 39, pp. 113-135.

Lusk J. L., Roosen J. and Fox J. A. (2003). ‘Demand for Beef from Cattle Administered Growth Hormones or Fed Genetically Modified Corn: A Comparison of Consumers in France, Germany, the United Kingdom, and the United States’, *American Journal of Agricultural Economics*, vol. 85, pp. 16-29.

Maaskant B., Jonkman S. N. and Bouwer L. M. (2009). ‘Future risk of flooding: an analysis of changes in potential loss of life in South Holland (The Netherlands)’, *Environmental Science & Policy*, vol. 12, pp. 157-169.

McFadden D. (1999). ‘Rationality for Economists?’, *Journal of Risk and Uncertainty*, vol. 19, pp. 73-105.

McFadden D. (2001). ‘Economic Choices’, *The American Economic Review*, vol. 91, pp. 351-378.

Meyerhoff, J. and Liebe, U. (2009) ‘Status Quo effect in choice experiments: empirical evidence on attitudes and choice task complexity’, *Land Economics*, vol. 85, 515-528.

Morkbak M., Christensen T. and Gyrd-Hansen D. (2010). ‘Choke Price Bias in Choice Experiments’, *Environmental and Resource Economics*, vol. 45, pp. 537-551.

Nagel H. and Hatzinger R. (1992). ‘Diagnostics in some discrete choice models’, *Computational Statistics & Data Analysis*, vol. 13, pp. 407-414.

NGENE (2010). NGENE 1.0.2 User manual & referefence guide. http://www.choice-metrics.com

Ohler T., Le A., Louviere J. and Swait J. (2000). ‘Attribute Range Effects in Binary Response Tasks’, *Marketing Letters*, vol. 11, pp. 249-260.

PBL (2010). *Evaluatie van een IPCC-klimaatrapport: een analyse van conclusies over de mogelijke gevolgen van klimaatverandering*, Planbureau voor de Leefomgeving, 500216001. (*in Dutch*)

Plott, C. R. (1996). ‘Rational individual behavior in markets and social choice processes: The discovered preference hypothesis,’ in *Rational foundations of economic behavior*, K. Arrow et al. (eds.), Macmillan and St. Martin's, London, pp. 225-250.

Racine J. S., Hart J. and Li Q. (2006). 'Testing the Significance of Categorical Predictor Variables in Nonparametric Regression Models', *Econometric Reviews*, vol. 25, pp. 523-544.

Rose J. M. and Bliemer M. C. J. (2009). 'Constructing Efficient Stated Choice Experimental Designs', *Transport Reviews: A Transnational Transdisciplinary Journal*, vol. 29, pp. 587-617.

Samuelson, W. and Zekchauser, R.J. (1988). 'Status Quo bias in decision making', *Journal of risk and uncertainty*, vol. 1, 7-59.

Scarpa, R., Ferrini, S., & Willis, K. (2005). 'Performance of Error Component Models for Status-Quo Effects in Choice Experiments,' in *Applications of Simulation Methods in Environmental and Resource Economics*, 6 edn, R. Scarpa & A. Alberini (eds), Springer Netherlands, pp. 247-273.

Slovic P. (1995). 'The construction of preference', *American Psychologist*, vol. 50, pp. 364-371.

Swait J. and Adamowicz W. (2001). 'The Influence of Task Complexity on Consumer Choice: A Latent Class Model of Decision Strategy Switching', *The Journal of Consumer Research*, vol. 28, pp. 135-148.

Swait J. and Louviere J. (1993). 'The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models', *Journal of Marketing Research (JMR).* vol. 30, pp. 305-314.

Train, K. E. (2009), *Discrete choice models with simulation*, 2nd Edition edn, Cambridge University Press.