UNIVERSITY OF GOTHENBURG
SCHOOL OF BUSINESS, ECONOMICS AND LAW

The first time is the hardest: A test of ordering effects in choice experiments

Fredrik Carlsson, Morten Raun Mørkbak, Søren Bøye Olsen

October 2010

EFMD

EQUIS
ACCREDITED

The first time is the hardest: A test of ordering effects in choice experiments

Fredrik Carlsson[a], Morten Raun Mørkbak[b], Søren Bøye Olsen[c]

Abstract

This paper addresses the issue of ordering effects in choice experiments, and in particular how learning processes potentially affect respondents' stated preferences in a sequence of choice sets. In a case study concerning food quality attributes of chicken breast filets, we find evidence of ordering effects in a sequence of 16 choice sets, where the last 8 choice sets are identical to the first 8. The overall preference structure is found to differ significantly between the two identical sequences of choice sets, and significant increases in marginal WTP are found for two out of four attributes. We find a reduction in the error variance for the last 8 choice sets relative to the first 8 choice sets. In particular, this difference is ascribed to the first choice set obtaining a significantly higher error variance than all succeeding choice sets, suggesting institutional learning rather than preference learning effects underlying the observed ordering effect. This is further supported by the fact that the differences in WTP become insignificant when removing the first choice set from the analysis. We find no evidence of fatigue, and we argue that our findings cannot be explained by starting point or strategic behavior effects.

Keywords: Choice Experiments, Fatigue, Learning, Ordering Effects, WTP

1. Introduction

Ample experimental evidence in stated preference surveys suggests that people do not make coherent choices, i.e., they are affected for instance by the context and various cues. This is clearly at odds with standard assumptions in economics, and also with the assumptions we make when analyzing stated preference (SP) responses. In this paper, we are particularly interested in investigating how learning effects potentially affect respondents' choices through a sequence of choice sets. This issue is related to what is known as ordering effects. Day et al. (2010) provide an excellent discussion of different explanations and manifestations of ordering effects. They discuss six different effects that are not necessarily mutually exclusive. The first one is preference learning or the discovered preference hypothesis, which relates to value uncertainty (Plott 1996). The hypothesis states that when respondents are faced with new decisions in unfamiliar environments, initial decisions will be incoherent and exhibit significant randomness. However, as choices are repeated and respondents gain familiarity with the decision environment, the decisions progressively become more coherent and less random. The second effect is referred to as institutional learning, which relates to the fact that most respondents participating in SP surveys have never experienced this type of survey before. Both preference learning and institutional learning suggest that one way to reduce uncertainty in SP surveys is to have respondents make repeated choices (Braga and Starmer 2005). Yet, the third effect discussed by Day et al. (2010) is fatigue. Respondents could get tired of the choice task if it is repeated many times, and thus, their choices may exhibit increasing levels of randomness over the sequence of choice tasks (Swait and Adamowicz 2001). The fourth effect potentially causing ordering effects is the starting point effect, where respondents who are uncertain about their preferences for the good regard a presented price as a cue to the "correct" value for that good, and consequently they anchor their WTP to this value (Kahneman et al. 1982). Finally, the fifth and the sixth effects are that respondents may

act strategically (Carson and Groves 2007; Day et al. 2010; Day and Pinto 2010).[1] In the case of provision of private goods, the effect of strategic behavior is not clear-cut since it depends on what assumptions one makes about the respondent's cognitive capacity. However, one typical behavior could be that respondents reject alternatives if they have already had an opportunity to obtain a similar alternative at a lower cost in a previous choice set. It should also be noted that an order effect may be counteracted by respondents' desire to act in a coherent way. This is what Ariely et al. (2003) call coherent arbitrariness, i.e., an individual's choices can be internally coherent, yet at the same time they can be anchored to the first choice or some initial starting point.

In the present paper we investigate to what extent preferences are stable in a choice experiment (CE) concerning food safety attributes of chicken breast filets. However, instead of varying the position of a given choice set, we apply an experimental setup using 16 choice sets per respondent where the last sequence of 8 choice sets is an exact copy of the first sequence of 8 choice sets. We estimate separate models based on the two sequences. The null hypothesis is that the preferences and the error variance do not differ between the two sequences.

Preference stability in stated preference surveys with repeated questions has received considerable attention in the literature. However, perhaps as could be expected, no consistent pattern has emerged. Furthermore, the majority of the studies have focused on and examined learning and fatigue effects by comparing the choices made in identical choice sets presented at different positions in a sequence of choice sets. A number of studies find no or small differences in preferences based on choices made in the beginning and choices made at the

---

[1] Day et al. (2010) distinguish between strategic behavior that is based upon a full recall of the presented prices (strategic behavior) and a recall that is weighted toward the recent prices seen (reference price behavior).

end of a sequence (see, e.g., Carlsson and Martinsson 2001; Johnson and Bingham 2001; Hanley et al. 2002; Brouwer et al. 2010). Other studies find that stated preferences do depend on the sequence of preference-eliciting choice questions. For example, Bateman et al. (2008b) test this by repeating the first choice set at the end of the sequence of choice sets. They find that respondents are less likely to choose an identical alternative when it is placed toward the end of a choice sequence and furthermore that the effect on the marginal utility of a given alternative is lower when placed toward the end of a sequence. It has also been found that respondents may suffer from starting point bias, or coherent arbitrariness, but that learning seems to reduce this bias (see, e.g., Ladenburg and Olsen 2008). A number of studies also show that a respondent's consistency in choices depends on the complexity of the task, on the positions of the choice sets in the sequence, and on his/her cognitive capability (see, e.g., Dellaert et al. 1999; DeShazo and Fermo 2002; DeSarbo et al. 2004; Lagerkvist et al. 2006; Brown et al. 2008; Savage and Waldman 2008; Day et al. 2010; Day and Pinto 2010).

Our design offers some additional insights compared to previous studies. First of all, by repeating exactly the same choice sets, the influence of exact sequencing of the choice sets is limited. Secondly, we can actually obtain the full set of preferences based on the two sequences, which facilitates a much more detailed comparison. Though our experimental design does not allow us to fully discriminate between all the different potential effects of ordering mentioned above, we argue that our results provide evidence mainly for learning effects taking place, and in particular institutional learning. We find that the overall preference structure differs between the two sequences of choice sets and that the error variance for the last sequence of 8 choice sets relative to the first sequence of 8 choice sets is significantly reduced. The willingness to pay (WTP) is higher for two of the attributes based on the responses of the last sequence. We find that it is primarily in the first choice set that the

error variance is high (compared with the other sets) and that the largest share of respondents change their preferences compared with an identical choice set given later in the sequence of choice sets. This suggests that it is primarily institutional learning and not so much preference learning that drives the observed ordering effect. While our results imply that learning effects can indeed be of significant structural importance when conducting CE surveys, our results also underline that if the focus is solely on obtaining policy advice in terms of WTP estimates, learning effects are not negligible.

The paper is organized as follows. Section 2 describes the econometric approach and Section 3 the design of the survey and the data. Section 4 presents the results and a discussion and Section 5 concludes the paper.

2. Econometric approach

The experiment consists of 16 choice sets, where the same 8 sets are given in two sequences. The last sequence (B) of 8 sets is an exact copy of the first sequence (A) of 8 sets. In the analysis, we apply a standard random utility model (McFadden 1974), where the utility of alternative j for individual i in choice set k in sequence S is specified as

$$V_{ijk}^S = v_{ijk}^S + \varepsilon_{ijk}^S = \beta_i^S a_{jk} + \varepsilon_{ijk}^S, \tag{1}$$

where a is a vector of attributes, β is the corresponding parameter, and ε_{ijk} is an error term. If the error terms are iid extreme value distributed with variance $\pi^2/(6\mu^2)$, the standard logit model choice probability that individual i chooses alternative j is

$$P_{ijk}^S = \frac{\exp{(\mu^S v_{ijk}^S)}}{\sum_{m \in k} \exp{(\mu^S v_{im}^S)}}, \tag{2}$$

where μ is a scale parameter that is inversely proportional to the error variance. The coefficients (β) in the econometric models are usually expressed in their scaled form *($\beta = \mu\beta^*$),* where the scale parameter μ and the "original" coefficients β^* cannot be separately identified due to confounding. Hence, the estimated parameter β indicates the effect of each observed variable relative to the variance of the unobserved factors (Train 2003).

Comparing the estimated models from Sequences A and B, there are thus two elements involved: the parameter vector and the scale parameters. The problem, though, is that the scaling factor cannot be identified in any particular set of empirical data. Yet, instead, the ratio of the scale factor of one data set relative to another can be identified by normalizing one of them to the value of 1 and then defining a range of values of the other scale factor, within which we expect the log likelihood function to be maximized (Swait and Louviere 1993). We draw on the notion used by, e.g., Holmes and Boyle (2005) and Savage and Waldman (2008) that learning processes can lead to increased consistency in choice, which implies reduced error variance and a higher degree of estimation precision (Heiner 1983; de Palma 1994; Hole 2007), whereas fatigue effects will have the opposite implication. These effects are of course in addition to any potential effects of learning or fatigue on the attribute parameter estimates.

In the analysis, we estimate random parameter models where we assume that all non-price attributes are normally distributed, thereby allowing consumers to place positive as well as negative values on the non-price attributes and the alternative specific constant, which has been shown to work well in earlier studies. The price coefficient is held constant since such an assumption of a constant price parameter allows straightforward calculation of the distribution of WTP. We have used the software package Biogeme (Bierlaire 2003) to estimate the models, allowing for a direct estimation of the ratio of scale factors. The models

are estimated with simulated maximum likelihood using Halton draws with 300 replications; see Train (2003) for details on simulated maximum likelihood and Halton draws. Specifically, we compare a model based on the first sequence of 8 sets with a model based on the second sequence of 8 sets, and this test could easily be conducted on larger number of models within the 16 choice sets (see below for further details on the design). In addition to comparing the estimated parameters and relative scale parameters between models, we also compute and compare marginal WTPs for each attribute. The advantage of such a comparison is that the scale parameters cancel from the expression and we can thus directly compare WTP estimates between the models.

3. The choice experiment

The choice experiment concerned food safety attributes of chicken breast filets. Prior to the design of the CE, we performed three focus group interviews. In the focus groups, the following attributes were identified as being important in relation to the choice of chicken breast filets: type of production, country of origin, and, to some extent, food safety (mainly related to Salmonella). Although food safety did not appear to be of great concern to consumers, the original purpose of the present study was to elicit the relative weighting of food safety. Consequently, we included two food safety attributes associated with the chicken products: "Salmonella-free" and "Campylobacter-free." These attributes were chosen because of their relevance to each of the products and also because they were judged as representing an increasingly important issue from a scientific as well as a political perspective. Table 1 presents the attributes and their associated levels.

Table 1: The attributes and their levels in the choice experiment for the chicken breast filet.

Attributes	Levels
Type of production	Conventional (indoor), organic (outdoor)
Country of origin	Domestic (Danish), non-domestic
Campylobacter-free	Not labeled Campylobacter-free, Campylobacter-free
Salmonella-free	Not labeled Salmonella-free, Salmonella-free
Price (DKK)	25, 28, 33, 40, 50, 65, 85, 115

Note: DKK 10 ~ EUR 1.34.

The two food safety attributes in the survey, Salmonella-free and Campylobacter-free, are quite similar. They both exhibit private good characteristics to a large extent, and both give rise to more or less the same course of illness. The main difference is that the current risk of getting infected by Campylobacter is much higher than the risk of contracting a Salmonella infection. Consequently, our *a priori* expectation of the value of a Salmonella-free chicken is that respondents will not value it as highly as the Campylobacter-free characteristic.

A D-efficient fractional factorial design resulting in a sequence of 8 sets was used for the experimental design. At the end of this sequence, the entire sequence of 8 sets was repeated, resulting in a total of 16 choice sets per respondent. Hence the fractional factorial design of 8 sets was presented twice to each respondent. Respondents were not made explicitly aware of this feature of the design and we had no indication that any respondent realized it. In each choice set, the respondents were faced with two alternative chicken breast filet products plus a third status quo alternative (all specified as packages of 500 grams). The latter characterized the respondents' usual purchase, which was identified earlier in the questionnaire. This approach of using the respondents' "own" status quo values has been recommended and used in other studies to mimic the actual purchasing situation as closely as possible (Ruby et al. 1998; Kontoleon and Yabe 2003). The use of individual specific status quo alternatives in the

design procedure of CE has been further developed by Rose et al. (2008), in what they refer to as segment-specific efficient designs, two-stage process designs, and individual efficient designs.

4. Results and discussion

The CE survey was conducted using an internet panel, and the sample was obtained from Nielsen's online database. In Denmark, there are approximately 2.4 million private households, of which 87% have access to the internet. All panel members are 15 years old or older and reside in a household with internet access, yet in the present survey only respondents above age 18 were allowed to participate. The final sample consisted of 389 respondents, each answering 16 choice sets, resulting in a total of 6,224 choice observations. The response rate was 26%. To begin with, we estimate the results based on the two identical sequences (i) the first 8 choice sets and (ii) the last 8 choice sets. Next, we pool the responses from the two sequences and estimate two additional models (iii) using all 16 choice sets, but not accounting for potential difference in scale between the first and last 8 sets, and (iv) using all 16 sets, allowing and accounting for potential difference in scale between the first and last 8 sets. Table 2 displays the results obtained in the four different models.

Table 2: RPL models for the first 8 and last 8 sets, and for pooled models; standard errors in parentheses.

Parameter estimates	Model (i) First 8 sets Coeff. (std. err.)	t-value	Model (ii) Last 8 sets Coeff. (std. err.)	t-value	Model (iii) All 16 sets - not corrected for scale Coeff. (std. err.)	t-value	Model (iv) All 16 sets - corrected for scale Coeff. (std. err.)	t-value
Campylobacter-free label	0.81 (0.067)	12.13	0.916 (0.089)	10.35	0.848 (0.054)	15.82	0.29 (0.020)	14.27
Salmonella-free label	0.62 (0.081)	7.66	0.95 (0.122)	7.76	0.778 (0.068)	11.46	0.27 (0.024)	11.39
Domestic produce	1.192 (0.099)	12.01	1.338 (0.160)	8.34	1.242 (0.084)	14.82	0.428 (0.031)	13.9
Outdoor production	0.454 (0.095)	4.77	0.764 (0.138)	5.55	0.52 (0.070)	7.48	0.181 (0.025)	7.35
ASC (Status quo)	0.142 (0.13)	1.09	0.32 (0.139)	2.29	0.277 (0.093)	2.96	0.0916 (0.032)	2.83
Price	-0.037 (0.004)	-14.07	-0.0441 (0.004)	-11.96	-0.041 (0.002)	-18.67	-0.014 (0.001)	-16.74
Standard Deviation estimates								
Campylobacter-free label	0.075 (0.212)	0.35	0.22 (0.29)	0.76	0.194 (0.103)	1.89	0.073 (0.035)	2.08
Salmonella-free label	0.54 (0.1404)	-3.84	0.84 (0.196)	-4.29	0.702 (0.115)	6.1	0.248 (0.037)	6.65
Domestic produce	0.676 (0.1496)	-4.53	1.038 (0.176)	-5.9	0.832 (0.114)	7.33	0.294 (0.038)	7.79
Outdoor production	1.046 (0.193)	5.42	1.236 (0.224)	-5.52	1.156 (0.131)	8.8	0.402 (0.045)	8.99
ASC (Status quo)	1.93 (0.138)	13.93	2.11 (0.193)	10.95	2.04 (0.112)	18.28	0.701 (0.043)	16.16
μ (first 8 CS = 1)							1.15 (0.08)	1.88 [a]
LL	-2534		-2415		-4969		-4966	
Adj. Rho-square	0.256		0.291		0.272		0.272	

[a] The t-value for the scale factor is a t-value tested against the null hypothesis H_0: $\mu = 1$.

Nearly all attribute coefficients are statistically significant at the standard 5% level, with the exception of the ASC in model (i), the standard deviation of the Campylobacter-free label in models (i) and (ii), and the scale factor in model (iv).

The difference in responses between the two sequences is initially examined through a likelihood ratio test for equality of all model parameters, including the scale parameters (Swait and Louviere 1993). This test involves Models (i) - (iii). The likelihood ratio is 40.29, which means that we can reject the hypothesis of equal parameters at the standard 5% level of statistical significance (critical value at 5% is 18.31). We therefore proceed with the pooled model where we allow for a difference in scale parameters between the two sequences. The likelihood ratio is 34.23, which means that we can reject the hypothesis of equal parameters at the standard 5% level of statistical significance as well (critical value at 5% is 16.92). Recall that the scale parameter is the inverse of the standard deviation of the error term in our specification (Swait and Louviere 1993). The estimated relative scale factor of 1.15 in Table 2 implies that the variance of the error term or "noise" in the model based on the last sequence is 76% of the variance of the model based on the first sequence.[2] This is also in accordance with the comparison of model fit between Models (i) and (ii), where the model based on the last sequence clearly provides a better fit to the data. The finding of a reduced error variance in the last sequence is equivalent to the findings by Holmes and Boyle (2005). Consequently, we observe a shift in the preferences for the attributes from the first to the second sequence and smaller error variance in the second sequence. This suggests that there is learning in the choice experiment, both in terms of changes in preferences and less noise. Do note that any attempt of respondents to be coherent goes in the opposite direction as learning.

The above tests are joint tests that demonstrate a change in preferences. In a more detailed analysis of *how* the preferences change, we estimate the unconditional marginal WTP for each attribute and the ASC, i.e., the ratio between the attribute coefficient and the price coefficient. Standard errors are estimated with the Delta method. Table 3 presents the results.

---

[2] $(1/1.15^2) = 0.7561$. Hence, there is a 24% difference in the variance between the two models.

Table 3: WTP estimates and t-values based on Models (i) and (ii); standard errors in parentheses.

Mean	Model (i): First 8 sets WTP	Model (ii): Last 8 sets WTP	H_o: WTP(i) = WTP (ii) t-value
Campylobacter-free label	21.60 (1.96)	20.77 (1.98)	0.418
Salmonella-free label	16.53 (2.01)	21.54 (2.65)	2.318
Domestic produce	31.79 (3.37)	30.34 (4.98)	0.500
Organic production	12.11 (4.11)	17.32 (5.28)	1.703
ASC (Status quo)	3.79 (12.57)	7.26 (11.12)	0.713
Standard deviation			
Campylobacter-free label	1.99 (15.97)	4.99 (20.75)	0.495
Salmonella-free label	14.40 (8.71)	19.05 (14.51)	0.964
Domestic produce	18.03 (10.10)	23.54 (13.04)	1.146
Outdoor production	27.89 (12.86)	28.03 (20.39)	0.023
ASC (Status quo)	51.47 (21.66)	47.85 (20.97)	0.555

The WTP estimates reported in Table 4 reveal that the differences in preferences revealed by the likelihood ratio test are mainly due to differences in mean parameters estimates for two of the attributes. The only significant differences in WTP between the two models are for the salmonella-free attribute and the value of organic production (at the 10% significance level). In both cases, the estimated WTP is higher in the second sequence. If the preferences stated in the second sequence are a better reflection of respondent preferences, then the responses from the first sequence will significantly underestimate the value of two of the quality attributes that we are interested in. Note that the shift in preferences is not driven by preferences for the status quo alternative as the WTP for the status-quo alternative is not significantly different between the two sequences. Our results are in contrast to a number of previous studies looking at the stability of preferences. However, many of the previous studies are between-sample tests (see, e.g., Carlsson and Martinsson 2001; Hanley et al. 2002).

Another interesting question is whether the change in preferences can be traced to certain choice sets or a certain part of the order of the choice sets. Given our within-subject design, we can make a number of comparisons. First of all we can look at the choices made in each choice set. Since we have two observations of choices for the same choice set, we can conduct symmetry tests, which are used to analyze matched-pair data with multiple discrete levels (see, e.g., Stata 2007). We can reject the hypothesis of no difference in responses between the two sequences for four out of eight choice sets at the 5% level. The pairs of choice sets where there is a significant difference are 1 and 9; 2 and 10; 5 and 13; and 8 and 16. However, it is for the first choice set that the largest difference in choices is observed. Comparing Set 1 with Set 9, almost 27% of the respondents change their answer, while for the other comparisons the share varies between 11 and 20%. This suggests that there is something particular with the first choice set. In order to further explore this, we re-estimate the pooled random parameter logit model, yet this time we estimate separate scale factors for Sets 2 to 16 relative to the normalized scale factor in Set 1, which is set to one. Table 4 presents the results.

Table 4: Estimation and comparison of choice set-specific scale factors in the pooled RPL model.

First 8 sets	1	2	3	4	5	6	7	8
μ	1	1.37	1.72	1.61	1.68	1.27	1.16	2.22
Std. Err.	0	0.17	0.20	0.21	0.23	0.17	0.23	0.23
t-value H_0: $\mu=1$		3.53	4.57	3.64	2.79	2.45	0.08	5.36
Last 8 sets	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
μ	1.54	1.65	1.91	1.80	1.60	1.46	0.98	1.98
Std. Err.	0.13	0.18	0.20	0.22	0.22	0.19	0.22	0.18
t-value H_0: $\mu=1$	2.18	3.69	2.92	2.94	1.58	0.70	5.33	4.32
t-value H_0: $\mu_{\text{First 8}} = \mu_{\text{Last 8}}$	4.32	1.12	0.68	0.63	0.25	0.75	0.56	0.82

Note: The vertical alignment of choice sets in the table corresponds to the choice sets that are identical.

As Table 4 shows, almost all of the estimated scale factors for Sets 2 to 16 differ significantly from the scale factor in the first set. Furthermore, it shows that they all are larger than one, and hence the error variance is smaller in these sets relative to the first set. This is further evidence that the behavior in the first choice set is different from the behavior in the other sets and that the noise is significantly larger in the first set. Moreover, there is no significant difference in scale factors between the pairs of choice sets other than between Sets 1 and 9. This is in contrast to the findings by Bateman et al. (2008b), where the most reliable estimates are found in the first 4 and last 4 sets, whereas the middle 8 sets somewhat surprisingly are the least reliable estimates. Our results are however largely in accordance with Bateman et al. (2008a), who found evidence of institutional learning particularly in the first valuation question, and Ladenburg and Olsen (2008), who found evidence that preferences become stable after the first three choice sets, supporting the notion of learning effects taking place. Similarly, both DeSarbo et al. (2004) and Brown et al. (2008) find evidence of learning effects in terms of increasing choice consistency as respondents progress through a sequence of choices.

If it is primarily the first choice set where the error variance is high (compared with the other sets) and where the largest share of respondents change their preferences compared with an identical choice set given later in the sequence of choice sets, then a comparison between the two sequences where the first set is dropped should result in smaller differences between the two sequences. In order to test this, we estimate a model based on Sets 2 to 9, i.e., Model (v), and compare this with the model based on Sets 9 to 16, i.e., Model (ii). There is thus one set overlapping, Set 9, which is necessary for identification of the main effects. Table 5 presents the resulting WTP estimates.

Table 5: WTP estimates and t-values based on Models (v) and (ii); standard errors in parentheses.

	Model (v): set 2-9 WTP	Model (ii): set 9-16 WTP	H$_o$: WTP(v) = WTP (ii) t-value
Mean			
Campylobacter-free label	25.55 (1.89)	20.77 (1.98)	0.902
Salmonella-free label	18.80 (1.93)	21.54 (2.65)	-1.281
Domestic produce	28.46 (3.14)	30.34 (4.98)	-0.659
Organic production	13.65 (3.87)	17.32 (5.28)	-1.213
ASC (Status quo)	3.99 (11.64)	7.26 (11.12)	-0.684
Standard deviation			
Campylobacter-free label	2.24 (23.09)	4.99 (20.75)	-0.416
Salmonella-free label	14.57 (6.33)	19.05 (14.51)	-0.981
Domestic produce	20.34 (6.22)	23.54 (13.04)	-0.729
Outdoor production	25.43 (11.08)	28.03 (20.39)	-0.462
ASC (Status quo)	50.48 (20.15)	47.85 (20.97)	0.411

As can be seen, there are no significant differences in WTP based on the responses to Sequences 2 to 9 and Sequences 9 to 16. Consequently, it is the behavior in the first choice set that explains the significant differences in WTP between the two sequences.

While we interpret our results as being in support of learning effects, there might be precedent-dependent effects that could explain why respondent choices differ in otherwise identical choice sets (Day and Pinto 2010; Day et al. 2010). Particularly, starting point effects and strategic behavior could affect respondent behavior. Although we have no possibility of formally testing the impact of such effects on our results, we would argue that the changes we find are mainly ascribed to position-dependent effects, particularly in terms of the institutional learning effect. Starting point effects could potentially explain the observed changes in marginal WTP estimates (Carlsson and Martinsson 2008; Ladenburg and Olsen 2008). Furthermore, the relatively high error variance in the first choice set might be regarded at least

as a precursor of starting point effects, bearing in mind the close relationship between preference uncertainty and starting point effects (Kahneman et al. 1982). However, the fact that the good being valued here is a very common market good, namely chicken breast filet with which the majority of the respondents have consumer experience, we would expect respondents to be relatively certain about their preferences (Plott 1996; List 2003). As a consequence, we argue that the relatively high error variance in the first choice set is mainly related to institutional uncertainty, while preference uncertainty is less likely. This is further supported by the fact that we see a very fast learning process since the error variance is significantly reduced already in the second choice set, which is indicative of institutional learning (Bateman et al. 2008a). Preference learning would arguably be a slower process as for instance in Ladenburg and Olsen (2008), where a starting point bias is found to persist for a number of choice sets. Thus, arguing that we see mainly institutional uncertainty in our data, we do not find it likely that our results are affected by starting point effects as we would only expect preference uncertainty, and not institutional uncertainty, to instigate such an effect. With respect to strategic behavior, Day et al. (2010) argue that marginal WTP decreases when respondents act strategically whereas Carson and Groves (2007) state that marginal WTP should not be affected by such behavior. The fact that we find an increase in marginal WTP for two out of four quality attributes is not in accordance with any of these explanations, suggesting that our results are not explained by strategic behavior.

5. Conclusions

This paper addresses the issues of how ordering effects related to the repeated choices in a choice experiment (CE) setting potentially affect respondents' stated preferences. The main focus is on changes in preference structure, error variance, and willingness-to-pay estimates. We use an an experimental design where a sequence of choice sets is given twice to the same

respondent. We find significant differences between the two sequences in both marginal WTP estimates and error variance. More specifically, the error variance is lower in the second sequence, suggesting that respondent uncertainty is lower. We find that it is primarily in the first choice set that the error variance is high (compared with the other choice sets) and that the largest share of respondents change their preferences compared with an identical choice set given later in the sequence of choice sets. This suggests that institutional learning is more likely to be present than preference learning. In this particular case study, mean WTP was found to be significantly higher for two of the attributes in the second sequence, but we have no reason to believe that this is a general result of institutional learning.

Overall, we find evidence in support of primarily institutional learning being the driver of the observed ordering effect rather than preference learning, fatigue, starting point bias, or strategic behavior. Our findings have clear implications for the design of choice experiments. One should be careful when including responses to a first choice set in the dataset used for estimating preference parameters and informing decision-makers. In fact, one should consider including an example of a choice set or an additional first choice set that is not generated by the statistical design and that is not intended for use in the analysis of preferences. This approach is likely to diminish any institutional uncertainty in CE and thus reduce potential ordering effects. It should however be noted that there is a risk of starting point bias if preference uncertainty is present, as may be the case especially in non-market good surveys. This suggests that the first choice set should be varied between survey versions in order to test and control for starting point bias.

References

Ariely D, Loewenstein G, Prelec D (2003) Coherent arbitrariness: Stable demand curves without stable preferences. Q J Econ. 118(1):73-105

Bateman I, Burgess D, Hutchinson W G, Matthews D I (2008a) Learning design contingent valuation (LDCV): NOAA guidelines, preference learning and coherent arbitrariness. J Environ Econ Manag 55(2):127-141

Bateman I, Carson R T, Day B, Dupont D, Louviere J, Morimoto S, Scarpa R, Wang P (2008b) Choice Set Awareness and Ordering Effects in Discrete Choice Experiments. CSERGE Working Paper EDM 08-01.

Bierlaire M (2003) BIOGEME: A free package for the estimation of discrete choice models. Proceedings of the 3rd Swiss Transportation Research Conference, Ascona, Switzerland.

Braga J, Starmer C (2005) Preference Anomalies, Preference Elicitation and the Discovered Preference Hypothesis. Environ Resource Econ 32(1):55-89

Brouwer R, Dekker T, Rolfe J, Windle J (2010) Choice Certainty and Consistency in Repeated Choice Experiments. Environ Resource Econ 46(1):93-109

Brown T, Kingsley D, Peterson G, Flores NE, Clarke A, Birjulin A (2008) Reliability of individual valuations of public and private goods: Choice consistency, response time, and preference refinement. J Publ Econ 92:1595–1606

Carlsson F, Martinsson P (2001) Do Hypothetical and Actual Marginal Willingness to Pay Differ in Choice Experiments? J Environ Econ Manag 41:179-192

Carlsson F, Martinsson P (2008) How much is too much?—An investigation of the effect of the number of choice sets, starting point and the choice bid vectors in choice experiments. Environ Resource Econ 40:165–176

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

Cummings R G, Brookshire D S, Schulze W D (1986) Valuing Environmental Goods: An Assessment of the Contingent Valuation Method. Rowman and Allenheld, Totowa, NJ

Day B, Bateman I, Carson R T, Scarpa R, Louviere J, Wang P, Dupont D (2010) Task Independence in Stated Preference Studies: A Test of Order Effect Explanations. Paper presented at the WCERE conference 2010, Montreal, Canada

Day B, Pinto JL (2010) Ordering anomalies in choice experiments. J Environ Econ Manag 59(3):271-285

de Palma A, Meyers G, Papageorgiou Y (1994) Rational choice under an imperfect ability to choose. Am Econ Rev 84:419-440

Dellaert B G C, Brazell J D, Louviere J (1999) The Effect of Attribute Variation on Consumer Choice Consistency. Market Lett 10(2):139-147

DeSarbo WS, Lehmann DR, Hollman FG (2004) Modeling dynamic effects in repeated-measures experiments involving preference/choice: an illustration involving stated preference analysis. Appl Psychol Meas 28:186-209.

DeShazo JR, Fermo G (2002) Designing Choice Sets for Stated Preference Methods: The Effects of Complexity on Choice Consistency. J Environ Econ Manag 44:123-143

Hanley N, Wright R E, Koop G (2002) Modelling Recreation Demand Using Choice Experiments: Climbing in Scotland. Environ Resource Econ 22:449-466

Heiner R (1983) The origin of predictive behavior. Am Econ Rev 73:560-595

Hole A R (2007) A comparison of approaches to estimating confidence intervals for willingness to pay measures. Health Econ 16:827-840

Holmes T, Boyle K J (2005) Learning and Context-Dependence in Sequential, Attribute-Based, Stated-Preference Valuation Questions. Land Econ 81(1):114-126

Johnson F R, Bingham M F (2001) Evaluating the validity of stated-preference estimates of health values. Swiss J Econ Stat 137:49-63

Kahneman D, Slovic P, Tversky A (1982) Judgement Under Uncertainty: Heuristics and Biases. New York: Cambridge University Press

Kontoleon, A. and Yabe, M. (2003) Assessing the Impacts of Alternative 'Opt-out' Formats in Choice Experiment Studies: Consumer Preferences for Genetically Modified Content and Production Information in Food. J Agr Pol Res 5:1-43.

Ladenburg, J. and Olsen, S. B. (2008) Gender-specific starting point bias in choice experiments: Evidence from an empirical study. J Environ Econ Manag 56(3):275-285

Lagerkvist CJ, Carlsson F, Viske D (2006) Swedish Consumer Preferences for Animal Welfare and Biotech: A Choice Experiment, AgBioForum 9(1):51-58

List J A (2003) Does Market Experience Eliminate Market Anomalies? Q J Econ 118(1):41-71

McFadden D (1974) Conditional Logit Analysis of Qualitative Choice Behavior. In: P. Zarembka (ed) Frontiers in Econometrics. Academic, New York

Plott C R (1996) Rational Individual Behaviour in Markets and Social Choice Processes: The Discovered Preference Hypothesis. In: K. Arrow et al. (ed) Rational Foundations of Economic Behaviour. Macmillan Press Ltd., London

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

Ruby M C, Johnson F R, Mathews K E (1998) Just say no: Opt-out alternatives and anglers' stated preferences. TER General Working Paper No.T-9801R.

Savage S J, Waldman D (2008) Learning and Fatigue During Choice Experiments: A Comparison of Online and Mail Survey Modes. J Appl Econometrics 23:351-371

Stata (2007) Stata Statistical Software: Release 10. College Station, TX: StataCorp LP.

Swait J, Adamowicz W (2001) The Influence of Task Complexity on Consumer Choice: A Latent Class Model of Decision Strategy Switching. J Consum Res 28:135-148

Swait J, Louviere J (1993) The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models. J Market Res 30:305-314.

Train K (2003) Discrete Choice Methods with Simulation. Cambridge University Press, Cambridge.