University of Warsaw
Faculty of Economic Sciences

Working Papers
No. 24/2014 (141)

Mikołaj Czajkowski
Anna Bartczak
Wiktor Budziński
Marek Giergiczny
Nick Hanley

Within- and between- sample tests of preference stability and willingness to pay for forest management

Warsaw 2014

University of Warsaw
Faculty of Economic Sciences

Within- and between- sample tests of preference stability and willingness to pay for forest management

Mikołaj Czajkowski
Faculty of Economic Sciences,
University of Warsaw
e-mail: miq@wne.uw.edu.pl

Anna Bartczak
Faculty of Economic Sciences,
University of Warsaw
Warsaw Ecological Economics Center
e-mail: bartczak@wne.uw.edu.pl

Wiktor Budziński
Faculty of Economic Sciences,
University of Warsaw
e-mail: wb311563@students.mimuw.edu.pl

Marek Giergiczny
Faculty of Economic Sciences,
University of Warsaw
Warsaw Ecological Economics Center
e-mail: mgiergiczny@wne.uw.edu.pl

Nick Hanley
University of St Andrews,
School of Geography and Sustainable Development
e-mail: ndh3@st-andrews.ac.uk

Abstract

The assumption of the stability of preferences is a fundamental one in the theory of the consumer. Many papers within the stated preferences literature have tested this assumption, and have found mixed results. Individuals may become more sure of their preferences as they repeat a valuation task or purchase decision; they may also learn more about prices and quantities of substitutes or complements over time, or about other relevant characteristics of both the good being valued and alternatives in their choice sets. In this paper, we test for the stability of preferences and willingness to pay for attributes of forest management both within and between samples. The within-sample test compares a set of responses from individuals over the sequence of a survey; the between-sample test compares responses from the same people over a period of 6 months. We find that respondents' preferences differ more within a sample (comparing their first 12 with their second 12 choices) than across samples. This may imply that preference learning and/or fatigue effects within choice experiments are more important than changes in preferences over time in this data.

Keywords:

preference stability, test-retest, discrete choice experiments, contingent valuation, stated preferences, forestry

JEL:

D01, H4, Q23, Q51

Acknowledgments:

This study was carried out as a part of the NEWFOREX project (New Ways to Value and Market Forest Externalities, FP7-KBBE-2009-3, Project no. 243950). The first author gratefully acknowledges the support of the Polish Ministry of Science and Higher Education and the Foundation for Polish Science.

1. Introduction

Brouwer (2012) notes that whilst "..in micro-economic theory, it is assumed that individuals know their preferences and *that these preferences are stable..",* the consensus from behavioural psychology is that individuals are continually (re-) constructing their preferences in a context-dependent manner. This implies that economists should expect that preferences for the same good, and WTP for a particular change in that good, might well vary over time for an individual, even if the time span over which preferences are observed is very short. Standard economic theory allows for WTP estimates to change as variables which co-determine one's demand for a good change, or as one learns more about one about the characteristics of a good (Munro et al., 2002) or one's preferences for experience goods (Czajkowski et al., forthcoming). However, preference parameters are supposed to be stable.

In this paper, we conduct both within- and between-sample tests of stability for choices, estimated preference parameters and willingness to pay. These tests are based on observations of the same individuals. The within-sample test considers responses to the first 12 and then second 12 choice questions in a survey on options for forest management. The between-sample test compares these choices with responses from an identical, 12-question choice experiment carried out six months later. This design provides a contrast with most of the test-retest literature, which as noted below has focussed on between-sample tests only. Our approach assists the analyst in determining why estimates of preferences and value from a stated preference exercise appear to change over time, and thus on the likely true extent of underlying instability of preferences for environmental goods. Within a sample of choice tasks, individuals may become more precise in stating their preferences, or may discover them as they gain experience in choosing between different bundles of a good. This perspective stands in contrast to between-sample tests, where an individuals' socio-economic conditions may change between samples, or where they may learn more about the good.

Within stated preferences, there have been a number of papers which also demonstrate a "time to think" effect on WTP for changes in an environmental good (Whittington et al., 1992; MacMillan et al., 2006). Specific tests for preference stability over environmental goods can be found in both contingent valuation and choice experiment settings. Using contingent valuation, Loomis (1989) applied test-retest procedures to compare estimates of Willingness to Pay (WTP) from two mail surveys which were conducted on the same respondents in nine months space. Surveys concerned the protection of Mono Lake in California. A comparison was done across two subsamples – the general population and visitors of Mono Lake. He applied t-tests to test the equality of average WTP between estimates from two surveys, based on a regression of WTP from first period on WTP from the second period, and comparing the estimated coefficients in these WTP models. Results indicate that preferences are stable over time between both subsamples. Carson et al. (1997) find similar results. They conducted two CV surveys

using a double-bounded elicitation format two years apart on different respondents in each period. The stability of preferences was investigated by testing the significance of interactions of explanatory variables with a dummy variable for the second period. Brouwer (2006) investigated the stability of preferences for public bathing water quality using two dichotomous choice surveys carried out over a nine month interval. Brouwer (2012) carries out a test-retest using CV with multiple payment formats (single-and double- bounded) for the protection of marine habitats in the Dutch North Sea. The re-test was undertaken with the same respondents 2 weeks after their original response, although here the response mode was changed from mail to telephone. Mean WTP fell slightly in the re-test sample, but not by a significant degree.

Test-retest procedures have also been applied to choice experiments in environmental economics. Bliem et al. (2012) estimate Multinomial and Mixed Logit on samples from two surveys of river restoration options in Australia, where the two surveys were undertaken one year apart. The model coefficients where compared using a Chow test. This indicated that there was no difference between preferences in these samples. Liebe et al. (2012) used an Error Component Logit model to compare preference and WTP estimates in two samples collected 11 months apart. Choices over on-shore wind power options were reasonably consistent over the interval, but WTP estimates differed significantly for around half of the attribute values. Schaafsma et al. (forthcoming) used a one-year interval to conduct a test-retest CE survey, and found that there were no significant changes in either preference parameters or WTP over this interval. However, the estimated error variance of choices fell over time. Most recently, Mørkbak et al. (2014) use CE to compare responses over a 2 week period for a market good (apples) with "real economic incentives". They thus seek to undertake a test-retest experiment in an incentive-compatible setting. They find "very good agreement" between the CE estimates of preferences over this rather short time interval. However, their sample consists of 25 persons only.

Further relevant contributions include Dupont et al. (2014), who compare estimates of WTP for health states related to water quality in Canada between surveys undertaken in 2004 and 2012, from contingent valuation and choice experiments. The health end points relate to stomach upsets from *e.coli* and bladder cancer. They found that whilst there was a significant change in WTP values across time when values were elicited using CV, there was no such significant change for the same values elicited using CE. A similar methodological comparison was undertaken by Brouwer et al. (2014), who survey the same sample of people in Switzerland at a 6-month interval using both CV and CE. Their study relates to willingness to pay for upgrading of waste water treatment plants in Switzerland to remove micro-pollutants. Some 20% of CE responses and some 30% of CV responses show no change in preferences over the 6 month interval. There was no significant difference, however, between WTP estimates over time for the sample as a whole, and no significant difference between CV and CE in this respect.

Beyond environmental applications, Ryan et al. (2006) investigated older peoples' preferences for social care. She uses the test-retest approach by resurveying the same respondents 11-60 days later. Coefficients of probit models with random effects estimated on samples from these two periods were insignificantly different, which may indicate stability of preferences, but it seems like these conclusions might be weak since only 47 respondents took part in second survey. Skjoldborg et al. (2009) conducted three DCE surveys on the same respondents over a period of 4 months to test preference stability for arthritis care programmes. A random effects logit model with interactions between attributes and two dummy variables indicating whether a response was from a second or third survey was estimated. The significance of these interactions was tested and WTP for each attribute in each period was calculated. Interactions and differences in WTP between periods were insignificant, which indicates stability of preferences.

Within-sample tests of preference stability within a choice experiment setting may reflect fatigue or learning effects. As people progress through a series of choice tasks, they may learn more what they like or do not like, so that they become more precise in their preferences, in the sense that their subjective distribution of their preference type becomes narrower as their experience in choosing increases (Czajkowski et al., forthcoming). Moreover, as people repeat choices, they may find that the choice task becomes simpler; or else they may become bored and start using heuristics more frequently (Swait et al., 2001). Any of these effects could show up as a change in the values implied by choices, whereas in fact there has been no shift in underlying "true" preferences. Such fatigue or learning effects could also show up in the random component of utility (Czajkowski et al., 2014).

2. Study Design and Data

2.1. The setting of the study – the Białowieża Forest

To apply the preference stability tests we used a Discrete Choice Experiment (DCE) study devoted to examining individuals' preferences and their willingness to pay for changes in the management of the Białowieża Forest in Poland. Such changes in management were aimed at (1) increasing the scale of "natural ecological processes" across different habitat types, and (2) restricting the number of visits to this popular forest. The Białowieża Forest is one of the most-widely recognised and ecologically valuable forests in Europe (Czajkowski et al., 2009). A considerable part of the Białowieża Forest (about 35% of the current forest area under National Park designation and in nature reserves) has remained practically un-affected by human activity for hundreds of years, due to the fact that this forest originally belonged to the kings of Poland and was used by them as hunting grounds where the removal

of wood was prohibited. This part of the Białowieża Forest is characterized by a high level of naturalness generally not present in other Polish and European forests.[3]

The entire area of the Bialowieża Forest cannot, however, be considered as the forest with the high level of naturalness. Most of the Białowieża Forest has been subjected to human activity and commercial use. Some 50% of its territorial area is similar to a typical commercial forest where management is focused on sustainable timber production. The forest also contains second-growth areas. During and immediately after the First World War, about 15% of the Bialowieża Forest was clear cut. This area was not re-planted, and the forest which currently grows there is a product of natural regeneration. The oldest trees there are around 80-90 years old. Although second-growth forests are less valuable than forests of the National Park and Nature Reserve areas, according to biologists they are still of conservation interest. Here, one can observe how forests emerge in places devoid of active management. According to the current law, however, this part of the forest can still be used for commercial purposes.

The Białowieża Forest is a popular destination for both Polish and foreign visitors. For several years, an increasing number of visitors to the Białowieża Forest has being noted, especially in summer and autumn during public holidays. Currently around 100,000 people visit the Białowieża Forest each year. During the public holidays in May and during the summer, the forest is visited by about 1,000 people each day. In the future, these numbers may increase. On the one hand, a large number of visitors may make the achieving of a higher level of forest naturalness difficult, and may even reduce the current level. Crowding can also adversely impact on utility per trip. For many years, environmentalists and various NGOs have been trying to convince decision makers to extend National Park designation to the entire Białowieża Forest; so far unsuccessfully, mainly due to local opposition, and to restrict visitor numbers. Our results can be thus used in public discussions on the enlargement of the Białowieża National Park and changes in visitor management.

2.2. Experimental design and data

The choice experiment comprised five attributes, of which one – forest naturalness in the National Park and natural reserves – could take only one level.[4] The other attributes were the future naturalness level of the current part of the Białowieża Forest under commercial management, the future level of

[3] Forests characterized by a high level of naturalness are understood as forests where: (1) ecological processes occur without human intervention, (2) there is a substantial share of trees older than 100 years, (3) the age of the trees varies naturally, and (4) there are large amounts of dead wood. All of these elements create favourable conditions for many rare species of animal, plant and fungus, which are not present in other forests.

[4] We decided to include this attribute in the choice sets to show respondents the entire picture of the current and future forest management of the Białowieża forest. Is very unlikely that in future the area currently under passive protection could be subject to less stringent protection

naturalness of the second-growth sections of the Białowieża Forest, the maximum visitor numbered allowed for the entire Białowieża Forest, and a cost. The payment vehicle was described as the increase in income taxes paid annually by each household which would pay for forest management. In the case of the naturalness for both parts of the Białowieża Forest, two management programs were considered, namely "commercial" or "passive" use, resulting in a low and high level of naturalness, respectively. The high level of naturalness[5] in the case of the second-growth forest was predicted to be reached in 150 years; whereas, the adequate time horizon for a commercial forest was defined as about 250 years.

Table 1. Attributes and levels used in the DCE study

Attributes	Levels	Variables
National Park and Natural Reserves (35% of the Białowieża forest)	High level of naturalness	*SQ*
Commercial forests (50% of the Białowieża forest)	Low level of naturalness	*SQ*
	High level of naturalness in 250 years	*COM*
Second-growth forests (15% of the Białowieża forest)	Low level of naturalness	*SQ*
	High level of naturalness in 150 years	*SGR*
Number of visitors	No restrictions	*SQ*
	5,000 visitors per day	*VIS1*
	7,500 visitors per day	*VIS2*
Annual cost per household	0 (*SQ*), 25, 50, 75, 100 PLN	*FEE*

The data used in this study was collected through a national online survey of the Polish population using two surveys carried out by the professional polling agency. The first survey was conducted in December 2011. Of the 1,443 respondents who agreed to take part in the web panel survey 1,000 successfully completed the survey resulting in response rate of 69.30%. The final sample was quota-controlled for sex, age, region and city size. In June 2012, each of the 1,000 participants who completed the first survey were invited to participate in a follow-up survey, and 789 of these agreed to participate. Since we were able to identify specific respondents within the data set, it provides an opportunity for us to test how their preferences changed over the course of the 6-month period, while controlling for the panel structure of the data and being able to compare the same individuals over time.[6]

[5] The high level of naturalness was the level of naturalness currently observed in the National Park and natural reserves. High level of naturalness was defined in terms of the following characteristics: (1) ecological processes occur without human intervention, (2) the forest is naturally composed with many tree species adjusted to local conditions, many of which are over 100 years old, (3) the age of the trees is naturally diverse and (4) trees are not cut – they remain in the forest until they grow old, die, fall and rot; there is plentiful of dead wood in such a forest (about 100 m^3/ha). This makes for favorable living conditions for many rare species of animals, plants and fungi – often not present in any other forests in Poland.

[6] We are making our data available upon request for replication of our results or any additional analyses.

The first survey included 24 choice tasks, which were in fact the same 12 choice tasks repeated twice (henceforth referred to as set A and set B, respectively). This design provides an additional opportunity for us to test preference stability on a choice-task by choice-task (within sample) basis. The second survey included 12 choice tasks (set C). The design of the study is illustrated in Figure 1. Comparing set A with set C for each person is thus a between-sample comparison of preference stability.

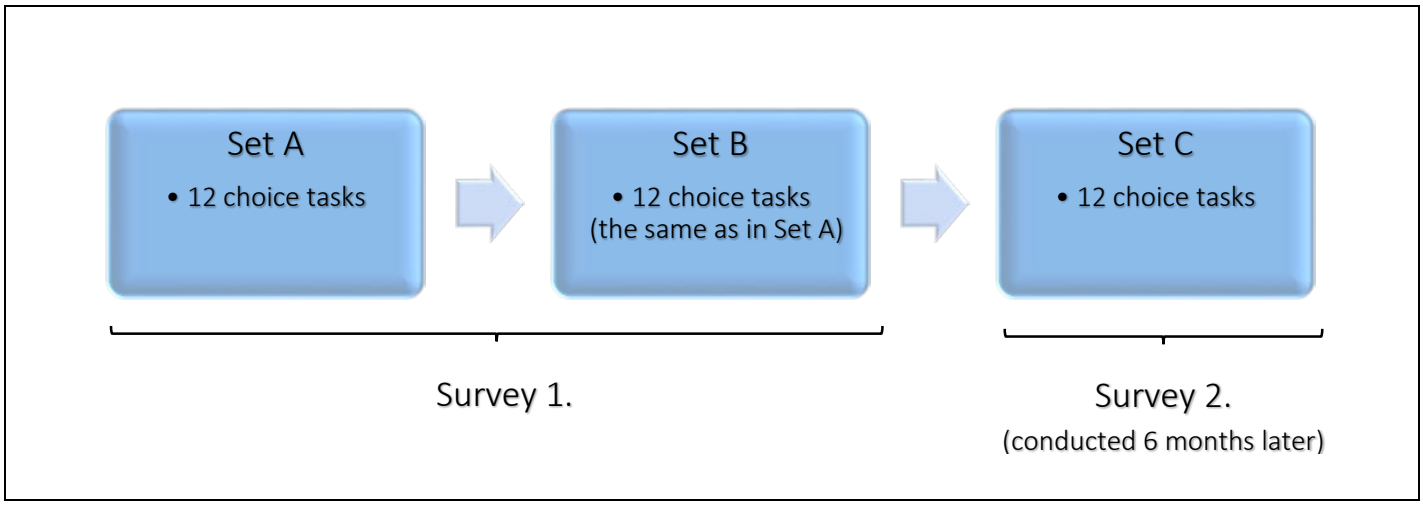

Figure 1. Illustration of the design of the preference stability study

The experimental design of our study utilized both an optimal-in-difference (Street et al., 2005; Street et al., 2007) and an efficient design (Ferrini et al., 2007; Scarpa et al., 2008) approach.[7] The designs were generated and applied separately for half of the participants of the first survey. In the second survey, half of the participants received the same optimal-in-difference design choice tasks, while the other half received updated efficient design choice tasks. This makes an additional comparison possible, for ¼ of our respondents (n = 193), to see how their choices changed between the first and the second survey, using exactly the same choice tasks.

Each choice task consisted of 3 alternatives, one of which was the SQ. In order to control for possible choice task- and position-specific ordering effects, each respondent was presented with a counterbalanced design in which (1) the order of choice tasks and (2) the order of alternatives (including the status quo alternative) was randomized, although it was kept constant for the same respondent in all

[7] The efficient design was generated using priors obtained from an MNL model estimated on the results of a pilot study conducted on a sample of 100 respondents. In order to account for uncertainty associated with our priors we employed Bayesian approach (Bliemer et al., 2008) in which we assumed all priors to be normally distributed with means estimated from the MNL model and standard deviations equal to 0.25 of each parameter mean.

choice tasks and surveys. We have taken steps to ensure that each choice task and each alternative was presented on every position a comparable number of times.

3. Non-parametric analysis

In what follows, two non-parametric tests of respondents' preference stability between sets A, B and C are presented. Because the choice tasks in sets A and B for each respondent were exactly the same, we are able to test if any respondents changed their answers and if so, then how and when. In addition, for the 193 respondents who were presented with the same choice tasks in sets A, B and C we are able to test if choices differed in any pair of sets (or a set and the remaining sets jointly). Finally, we are able to test if respondents' attitudinal answers changed between the first and the second survey.

3.1. Within-sample comparison

Figure 2 shows the distribution of the number of respondents who changed their choices between the sets A and B a particular number of times. Overall, only a minority of respondents (16.98%) made the same choice in every one of the repeated 12 choice tasks (and nearly 70 % of these always chose the SQ alternative). 59% of respondents changed decisions no more than 3 times, while 6.72% of respondents changed their decisions 8 times or more.[8]

[8] note that by making decisions completely randomly a respondent would, on average, make 8 different choices over the course of 12 choice tasks.

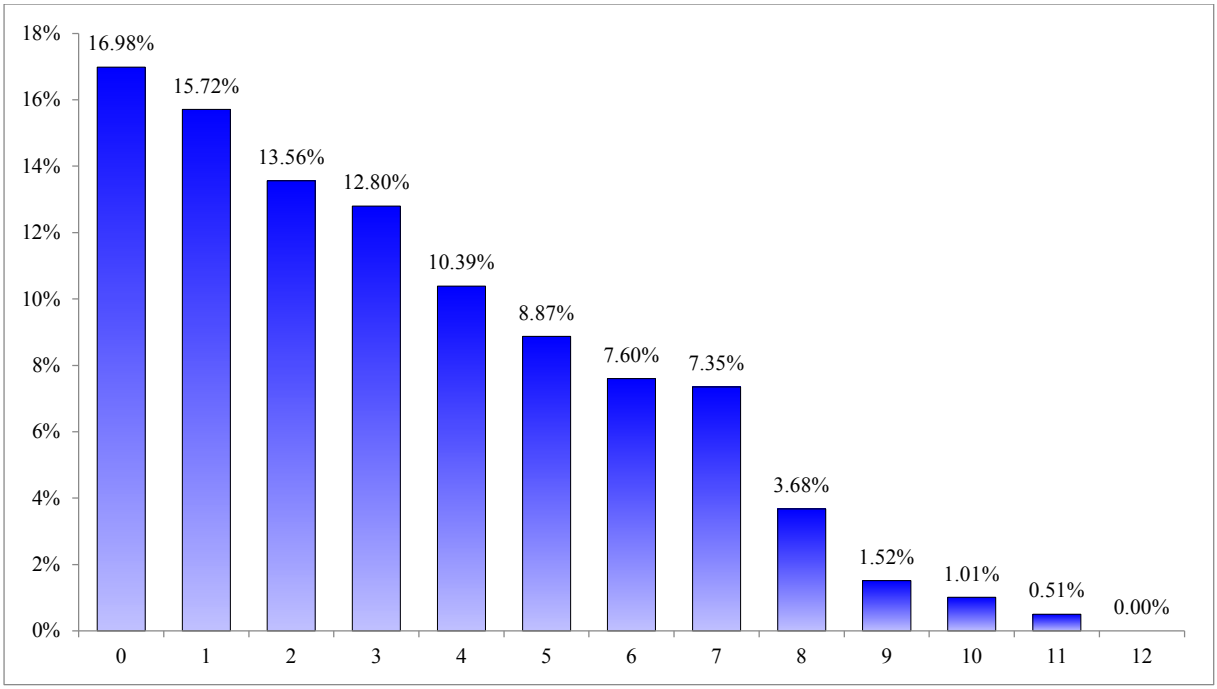

Figure 2. The distribution of respondents who changed their decisions a particular number of times in the series of 12 choice tasks repeated twice (sets A and B)

Additional insights are provided by the analysis of the dynamics of decision changes illustrated in Figure 2. In addition to the share of unchanged decisions in each choice task we present their (binomial proportion) confidence intervals, as well as sample means and their confidence intervals. No dynamics appear significant – the changes of decisions appear relatively uniformly distributed across choice tasks. This can be an indication of a preference stability combined with a relatively high degree of variability in respondents' choices (a low scale parameter, i.e. high utility function error term variance in relation to its deterministic part). If respondents' preferences change between choice tasks (e.g., due to preference learning), it does not appear to change their decisions in a significant way. In addition, we find no difference between respondents who chose the SQ and non-SQ alternatives – both groups seemed equally likely to change their decisions (possibly being an indication of existence of a group of respondents in a sample who made their decisions completely randomly, from the econometrician's perspective). However, respondents who chose non-status quo alternatives were less likely to be fully consistent in all 12 choice tasks – it seems easier for the respondents' who chose the SQ alternative in every choice to be consistent, as indicated by the shares of cumulatively unchanged decisions presented as bar plots in Figure 3.

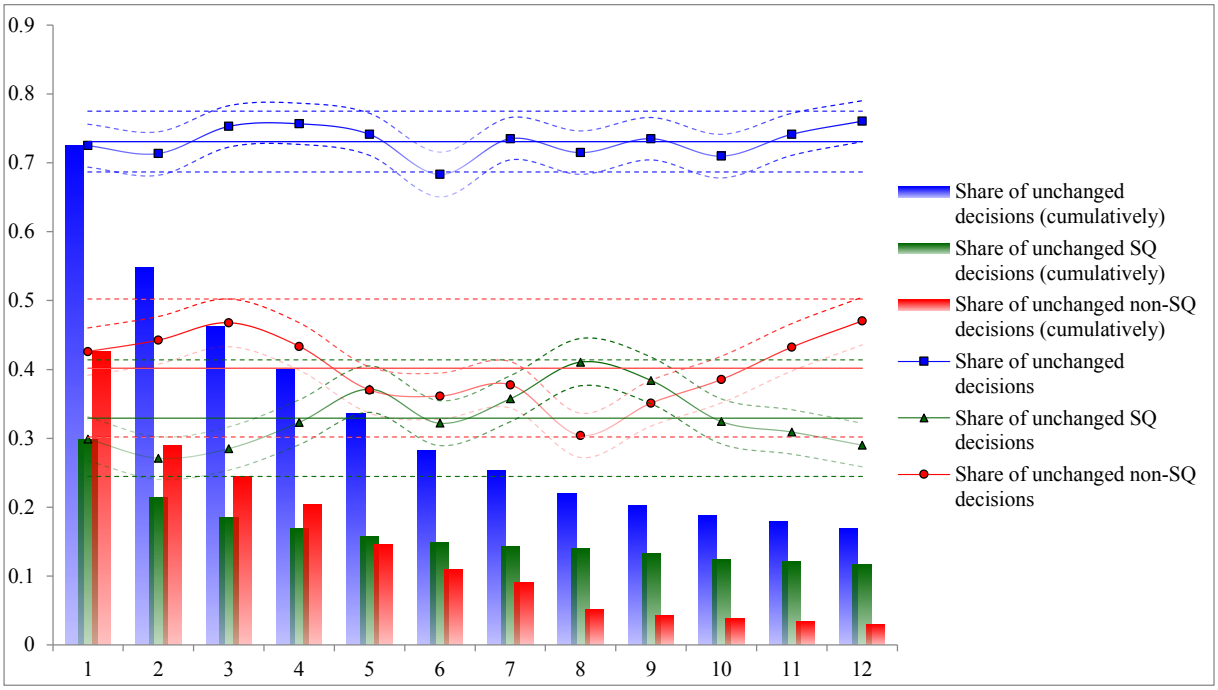

Figure 3. Dynamics of the number of decision changes in different choice tasks (set A vs. B; $n = 789$)

3.2. Between-sample differences

Figure 4 presents the dynamics of decision changes between surveys (i.e., between sets B and C). Overall, there is more variation in this case, but this is partly a result of a smaller sample, since in this case there were only 193 respondents who were presented with the same choices in both surveys. Once again, however, there are no significant changes in which choices (chronologically) are more likely to be answered in the same way as before. Interestingly, in this case the only respondents who answered in the same way in choice tasks are the ones who always chose the SQ alternative.

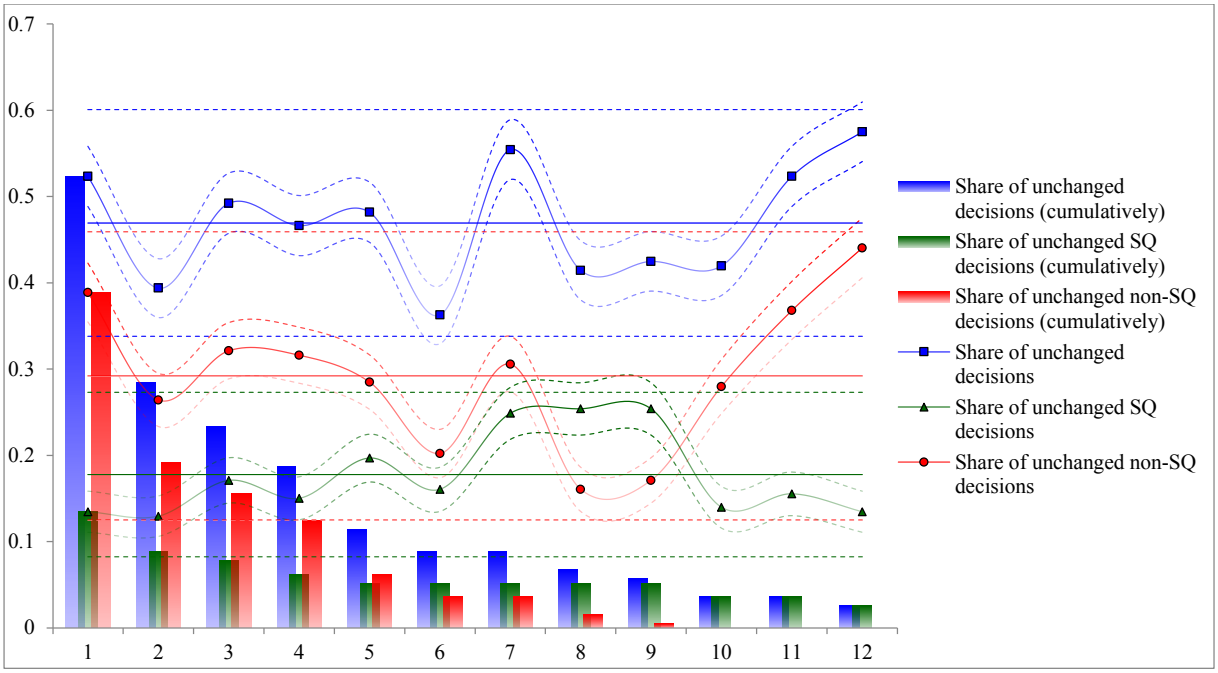

Figure 4. Dynamics of the number of decision changes in different choice tasks (set B vs. C; $n = 193$)

Finally, we compare measures of attitudes to forest management within and across the surveys. Annex A presents the attitudinal questions regarding the Białowieża Forest which were asked in both surveys. The Tables present a cross-tabulation of responses in the surveys 1 and 2. As well as the results of the Kruskal-Wallis test for the equality of the distributions of responses, we provide Wilcoxon signed rank tests for the equality of their medians. The results show that respondent's attitudes did not change significantly within or between surveys, except for the statement 'Białowieża Forest should be protected', for which their opinions became more diverse (although the mean response remained the same) and s shifts in responses to the question about "how should the National Park and Natural Reserves parts of Białowieża Forest be managed" (with an increase of respondents stating that it should be used commercially rather than passively protected, although the latter is still a much more popular opinion). For the other questions, however, attitudes remained unchanged.

4. Econometric analysis

This section presents the results of parametric analysis of respondents' choices based on the random utility framework (McFadden, 1974). We test if respondents' preferences were stable between groups of choice sets A and B presented in survey 1, and choice set C presented in survey 2 some six months later. We then estimate these models in WTP-space in order to investigate the changes in respondents' marginal willingness to pay for different attributes.

4.1. Stability of taste parameters

We start by investigating respondents' preference parameters stability, while taking preference and scale heterogeneity into account. To do this, we estimate conditional logit models in which all attribute-level related variables were interacted with binary variables for sets A, B and C. Evaluating the statistical significance of the differences of the estimates for different sets provides a convenient way of testing if respondents' preferences revealed in sets A, B and C are stable.

Table 2 presents the results of multinomial logit (MNL), and mixed logit (MXL, Revelt et al., 1998) models. In the MXL specification we assumed that all taste parameters are random and normally distributed. We accounted for the fact that each respondent faced 36 choice tasks (2 x 12 in the first and an additional 12 in the second survey) by allowing for individual-specific random coefficients and scale, where applicable. Estimation was performed using Matlab 8.1. The maximum likelihood function was simulated using 1000 shuffled Halton draws. Standard errors of coefficients associated with standard deviations of random parameters were simulated using Krinsky and Robb method with 10^6 draws (Krinsky et al., 1986).

The results presented in Table 2 show that, overall, respondents were in favor of extending passive protection to commercial (*COM*) and second-growth (*SGR*) areas of the Białowieża Forest, and disliked the status quo (negative but heterogeneous preference parameters associated with alternative specific constant for the no-change alternative). Preferences for restricting the number of visitors (*VIS1*, *VIS2*) were highly heterogeneous (coefficient of variation up to 5) which caused the coefficients of the means for random taste parameters to change signs between different model specifications.

Table 2. The results of conditional logit models of consumers' preferences accounting for preference differences between sets A, B and C.

	MNL			**MXL**					
	param. (s.e.)			**mean (s.e.)**			**s.dev. (s.e.)**		
	Set A (survey 1.)	**Set B (survey 1.)**	**Set C (survey 2.)**	**Set A (survey 1.)**	**Set B (survey 1.)**	**Set C (survey 2.)**	**Set A (survey 1.)**	**Set B (survey 1.)**	**Set C (survey 2.)**
SQ	0.1598***	0.3984***	0.1227**	-0.5318***	-0.9914***	-0.6780***	2.9109***	3.4200***	3.0198***
	(0.0537)	(0.0549)	(0.0538)	(0.0847)	(0.1018)	(0.0868)	(0.0848)	(0.1028)	(0.0829)
COM	0.4865***	0.5636***	0.5586***	0.5392***	0.7017***	0.6426***	0.6000***	1.0112***	0.6968***
	(0.0293)	(0.0295)	(0.0295)	(0.0389)	(0.0489)	(0.0356)	(0.0402)	(0.0458)	(0.0372)
SGR	0.6091***	0.6853***	0.5760***	0.7506***	0.9287***	0.7005***	0.6538***	1.0301***	0.5715***
	(0.0308)	(0.0308)	(0.0306)	(0.0439)	(0.0495)	(0.0342)	(0.0428)	(0.0490)	(0.0415)
VIS1	0.0681	0.0365	0.1354***	0.1102**	0.0575	0.1747***	0.0000	0.0000	0.0000
	(0.0401)	(0.0403)	(0.0401)	(0.0512)	(0.0533)	(0.0512)	(0.0726)	(0.0826)	(0.0759)
VIS2	0.1882***	0.2094***	0.2440***	0.2696***	0.3200***	0.3114***	0.0000	0.1124	0.2513***
	(0.0377)	(0.0374)	(0.0378)	(0.0465)	(0.0495)	(0.0411)	(0.0569)	(0.0663)	(0.0585)
FEE	0.0142***	0.0114***	0.0154***	0.0251***	0.0257***	0.0264***	0.0186***	0.0262***	0.0225***
(pref. space equivalent)	(0.0006)	(0.0006)	(0.0006)	(0.0010)	(0.0012)	(0.0010)	(0.0008)	(0.0009)	(0.0008)
Log-likelihood (constants only)		-30833.56				-30833.56			
Log-likelihood		-29135.74				-20452.24			
McFadden's pseudo R^2		0.0551				0.3367			
AIC/*n*		2.0528				1.4426			
n (observations)		28404				28404			
k (parameters)		18				36			

We find that the taste parameters appear to change from one set of choice tasks to another. In addition, estimates of standard deviations are usually higher for sets B and C than A. Nonetheless, it is difficult to draw specific conclusions without applying a quantitative tool to test if the changes are significant. To this end we applied the Wald test of the restriction that taste parameters (their distributions or only means of these distributions) are equal between different sets.[9] The results are presented in Table 3.

Table 3. Wald test statistics associated with restricting the taste parameters to be equal between sets of choices

	MNL		**MXL**			
	Equality of coefficients		**Equality of entire distribution**		**Equality of means**	
	Wald test statistic (d.f.)	p-value	Wald test statistic (d.f.)	p-value	Wald test statistic (d.f.)	p-value
A = B	21.7342*** (6)	0.0014	60.3145*** (12)	0.0000	25.627*** (6)	0.0003
A = C	8.9527 (6)	0.1763	120.6982*** (12)	0.0000	4.7494 (6)	0.5763
B = C	38.1744*** (6)	0.0000	178.6714*** (12)	0.0000	32.7918*** (6)	0.0000
A = B = C	45.6817*** (12)	0.0000	219.2699*** (24)	0.0000	39.6576*** (12)	0.0001

***, **, * – p-value lower than 0.01, 0.05, 0.10

The Wald test statistic is chi-squared distributed with the number of degrees of freedom equal to the number of restrictions. A high value of the statistic is an indication that the restriction significantly worsens the model and that the null hypothesis should be rejected. In our case, the overall hypothesis of preference stability (A = B = C) is rejected. We find that set A and set B shows significant differences, indicating that implied preference values changed over the sequence of choice tasks proceeding from the first 12 to the second 12 tasks within the first survey. A similar finding exists when comparing set B with set C: preferences implied in choices made 6 months later were statistically different from preference implied by choices in the second half of the first survey. Irrespectively of the model specification or sets for which the preferences are compared, we find that entire distributions, as well as their means, differ. Overall then, preferences do no t appear to be stable. Interestingly, however, differences between set A and set C are smaller (A and C are insignificantly different for the MNL model and for the means in MXL model). In other words, preferences revealed in set A are more similar

[9] It is important to note that the Wald test does not check if respondents changed their answers between samples, it tests if the distributions associated with taste parameters remained the same.

to set C than to set B. *Thus, the within-sample differences in preference estimates are greater for individuals than the between sample differences.*

4.2. WTP changes

In order to illustrate the implications of these preference changes we now turn to analyzing the changes in willingness to pay for forest attributes. By doing this, i.e. calculating marginal rates of substitution between attribute levels and the monetary attribute, we are able to cancel out proportional changes in all parameters which could be associated with the changes in utility function scale parameter – i.e., the 'error term' variance which is necessary from the econometrician's perspective, rather than with changes in taste parameters themselves. To do this, we estimate the same models, except we now do this in WTP-space (Train et al., 2005; Hensher et al., 2011).[10] The results are presented in Table 4.

Because the models were estimated in WTP-space, the parameters for all attributes but the cost can directly be interpreted as WTP (in PLN per year)[11]. For the model with random parameters we assumed that WTP for all attributes was normally distributed, while the negative cost parameter (*FEE*) was log-normal. The cost parameters are necessarily normalized to one in WTP-space models; instead we present their preference-space counterparts (coefficients of the underlying normal distribution).

A few interesting observations with respect to respondents' WTP dynamics can be made. First of all, we found that for the two nature protection related attributes (*COM* and *SGR*), respondents' WTP implied by their choices in set B is statistically higher than that resulting from choices in set A (their second set of 12 choices in survey 1 compared to their first set of 12 choices in survey 1). Interestingly, however, their choices conducted in the second survey (Set C), 6 months later, imply a decrease in WTP which becomes closer to WTP observed in Set A than to that observed in Set B. This finding is consistent with the results of the Wald test for equality of respondents' taste parameters' distributions noted above. A possible explanation of this is that as respondents proceed through many choice tasks they become less sensitive to the cost. After a few months, however, when they are presented with the survey and choice tasks again, their preferences appear closer to the preferences expressed in the beginning of the last survey, than at the end of that survey.[12] Another possibility is that they learn their preferences as

[10] As far as we know, this is the first application of the GMXL model in WTP-space in environmental and resource economics.

[11] 1 PLN ≈ 0.25 EUR ≈ 0.33 USD

[12] This explanation could also be interpreted in the light of the theoretical conditions of incentive compatibility proposed by Carson et al. (2007). According to them, stated preference survey should consists solely of one choice task. Several studies aimed at comparing a single binary question with a sequence of such questions were conducted in a field (Racevskis et al., 2008; McNair et al., 2011; Vossler et al., 2012; Scheufele et al., 2013), as well as in the lab (Bjornstad et al., 1997; Carson et al., 2009). The empirical examinations usually show

they move through the choice tasks of set 1, and then loose precision in their self-assessment of their preference type by the time set 2 data is collected six months later.

We also find that respondents' WTP for restricting the number of visitors to the Białowieża Forest were not statistically different between choice tasks. The WTP associated with the alternative specific constant representing the SQ alternative was more variable. This is not very surprising, however, considering the very high level of preference heterogeneity associated with this attribute (note the estimates of its standard deviations in different models). Once again, however, the WTP estimated from choices in Set A were closer to those from set C, than set B. The dynamics of WTP changes for different model specifications are presented in Figure 6.

inconsistency between value estimates following from a sequence of binary choice question and the single yes-no task, however, the evidence is not univocal. The theoretical analysis of Carson et al. (2007) suggests that the problem of the lack of incentive compatibility of formats consisting of several choice situations results from violation of the independence between choice sets, which is empirically proved in the laboratory experiment of Carson et al. (2009). On the other hand, Vossler et al. (2012) present contrary evidence suggesting that choice task independence is not important for truthful preference revelation.

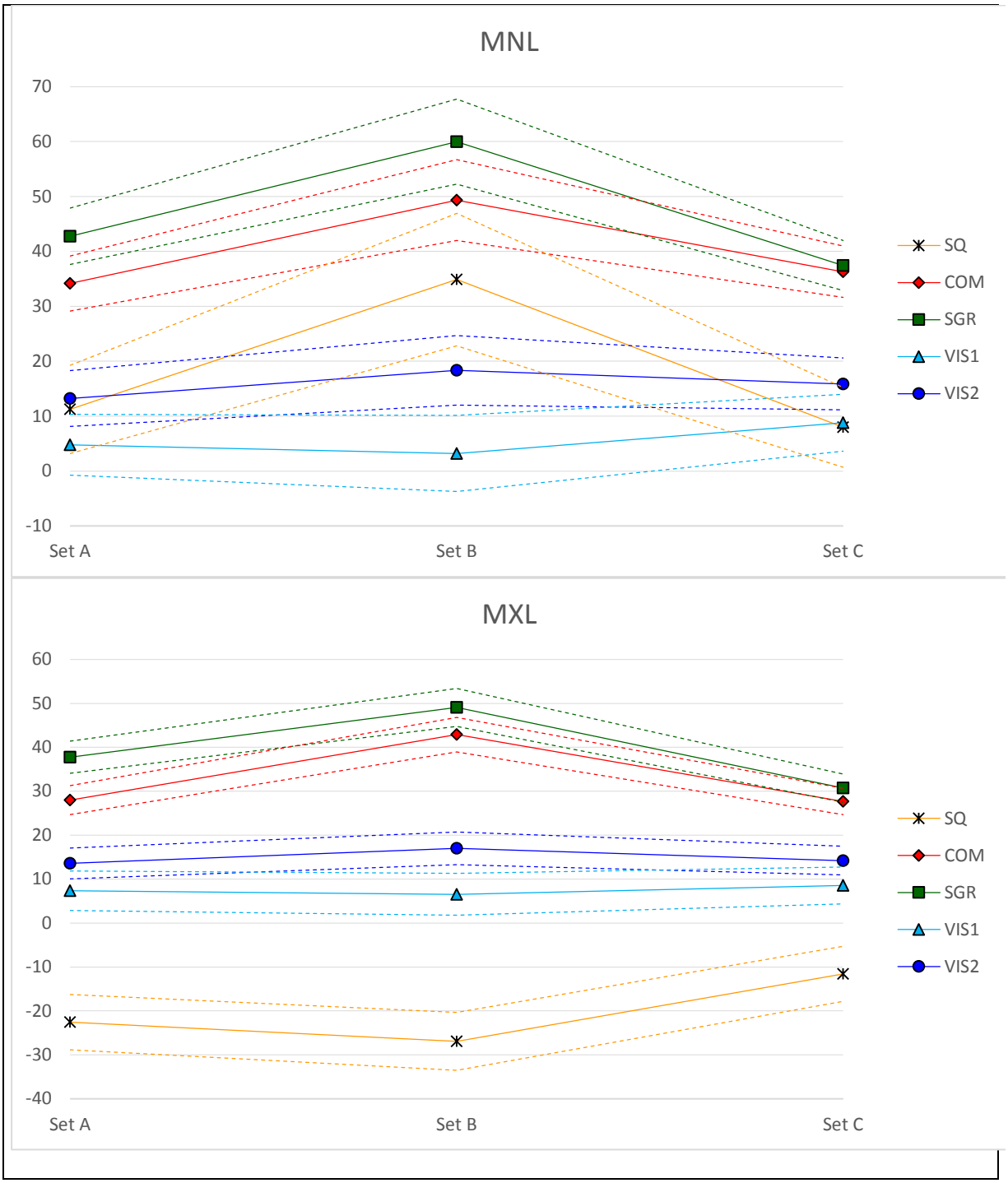

Figure 6. Dynamics of respondents' WTP changes between sets A, B and C in different models specifications

Table 4. The results of conditional logit models of consumers' WTP accounting for preference differences between sets A, B and C

	MNL			**MXL**					
	param. (s.e.)			**mean (s.e.)**			**s.dev. (s.e.)**		
	Set A (survey 1.)	**Set B (survey 1.)**	**Set C (survey 2.)**	**Set A (survey 1.)**	**Set B (survey 1.)**	**Set C (survey 2.)**	**Set A (survey 1.)**	**Set B (survey 1.)**	**Set C (survey 2.)**
SQ	11.2267***	34.8689***	7.9715**	-22.5485***	-26.9214***	-11.5583***	150.5604***	215.3832***	140.9194***
	(4.0952)	(6.1602)	(3.7026)	(3.2067)	(3.3651)	(3.2043)	(7.2115)	(11.191)	(5.6269)
COM	34.1528***	49.3278***	36.2908***	27.9639***	42.8771***	27.6774***	22.7431***	28.5611***	28.0879***
	(2.5506)	(3.7536)	(2.3782)	(1.683)	(2.0048)	(1.5216)	(1.7036)	(1.9887)	(1.6284)
SGR	42.7507***	59.972***	37.4215***	37.7418***	49.0675***	30.7018***	25.6269***	46.7267***	30.3228***
	(2.6238)	(3.9484)	(2.316)	(1.8665)	(2.2051)	(1.6343)	(1.8349)	(2.5352)	(1.8021)
VIS1	4.7874	3.1969	8.7975***	7.3629***	6.5275***	8.5553***	0.0000	0.2535	0.0000
	(2.8236)	(3.5335)	(2.6346)	(2.2893)	(2.4273)	(2.1491)	(2.6003)	(3.1658)	(2.8956)
VIS2	13.2147***	18.3311***	15.851***	13.5688***	16.9843***	14.1933***	0.0000	13.4896***	15.1765***
	(2.5969)	(3.2310)	(2.4098)	(1.7914)	(1.8941)	(1.6677)	(1.9550)	(2.3830)	(2.0412)
FEE	0.0142***	0.0114***	0.0154***	-3.7948***	-3.8118***	-3.7398***	0.5736***	0.8309***	0.5386***
(pref. space equivalent)	(0.0006)	(0.0006)	(0.0006)	(0.0430)	(0.0493)	(0.0340)	(0.0327)	(0.0385)	(0.0303)
Log-likelihood (constants only)	-30833.56			-30833.56					
Log-likelihood	-29135.74			-20669.04					
McFadden's pseudo R^2	0.0551			0.3297					
AIC/*n*	2.0528			1.4579					
n (observations)	28404			28404					
k (parameters)	18			36					

5. Conclusions

The temporal stability of preferences is a key "working hypothesis" in applied welfare economics and the theory of value. It is thus something which is important to test. For environmental goods, several such tests are reported in the literature using stated preference methods. Tests are typically made using samples from the sample individuals or repeated, non-identical samples from the same population over a time interval such as one month or one year. If we observe that preferences or WTP has changed, then this can be attributed to several factors which are hard to separately identify. Over time, people can change the relative importance they attach to goods and services, possibly through learning more about substitutes and complements for this good, and/or the characteristics of these goods. Incomes and other socio-economic attributes of individuals can also change over time, impacting on willingness to pay. Between-sample tests examine such changes.

But we can also look at how stated preferences and WTP evolve within an experiment if a choice modelling approach is taken, where people complete multiple choice tasks. Here, any differences in estimated preferences could be due to people learning their true preference type with more precision as they repeat the choice tasks, learning how to choose better with repetition, or becoming bored and thus falling back onto heuristics whilst making choices. Such effects are best studied in within-sample tests.

Our contribution is to combine within- and between-sample tests of preference stability for the same set of individuals. The within-sample test compares preferences (and implied willingness to pay) between the first 12 choices and the second set of 12 choices within the same survey. The between-sample test involved re-surveying the same individuals six months later, and repeating the same first 12 choice tasks. This allowed us to examine the dynamics of choices. Consumers often change their choices, but there is no obvious pattern in terms of *when* these changes occur. Deviations from previous choices were relatively uniformly distributed. Having said that, respondents who consistently choose the SQ alternative were less likely to change their choices than those who did not prefer the status quo.

Preferences differed between Sets A, B and C, but there were more significant differences in the within-sample comparison (set A compared to B) than the between-sample comparison (set B compared to C). On the whole, the differences between preferences in set A and set C were insignificant. The analysis of WTP changes shows a similar patter, but also that this effect is not the same for all attributes. For some attributes respondents are willing to pay more as they complete more choice tasks, but their WTP goes down again after 6 months. This results in the WTP values for set A being closer to set C than they are to set B. For the other attributes there are no significant differences in WTP between sets. Since the within-sample difference is much greater than the

between-sample difference. This suggests that the change in WTP in the first survey between the first 12 and second 12 choice tasks reflects something about the mechanics of choice, rather than a change in WTP as might be expected to occur over six months of time elapsed.

References

Bjornstad, D., Cummings, R. G., and Osborne, L., 1997. A Learning Design for Reducing Hypothetical Bias in the Contingent Valuation Method. *Environmental and Resource Economics*, 10(3):207-221.

Bliem, M., Getzner, M., and Rodiga-Laßnig, P., 2012. Temporal stability of individual preferences for river restoration in Austria using a choice experiment. *Journal of Environmental Management*, 103(0):65-73.

Bliemer, M. C. J., Rose, J. M., and Hess, S., 2008. Approximation of Bayesian Efficiency in Experimental Choice Designs. *Journal of Choice Modelling*, 1(1):98-127.

Brouwer, R., 2006. Do stated preference methods stand the test of time? A test of the stability of contingent values and models for health risks when facing an extreme event. *Ecological Economics*, 60(2):399-406.

Brouwer, R., 2012. Constructed preference stability: a test–retest. *Journal of Environmental Economics and Policy*, 1(1):70-84.

Brouwer, R., and Logar, I., 2014. Do choice experiments produce more stable welfare measures than contingent valuation? A test-retest. paper presented at the Fifth World Congress of Environmental and Resource Economists, 28 June - 2 July 2014, Istanbul, Turkey.

Carson, K. S., Chilton, S. M., and Hutchinson, G. W., 2009. Necessary Conditions for Demand Revelation in Double Referenda. *Journal of Environmental Economics and Management*, 57:219–225.

Carson, R., and Groves, T., 2007. Incentive and informational properties of preference questions. *Environmental and Resource Economics*, 37(1):181-210.

Carson, R. T., Hanemann, W. M., Kopp, R. J., Jon, A. K., Mitchell, R. C., Presser, S., Rudd, P. A., Smith, V. K., Conaway, M., and Martin, K., 1997. Temporal Reliability of Estimates from Contingent Valuation. *Land Economics*, 73(2):151-163.

Czajkowski, M., Buszko-Briggs, M., and Hanley, N., 2009. Valuing Changes in Forest Biodiversity. *Ecological Economics*, 68(12):2910–2917.

Czajkowski, M., Giergiczny, M., and Greene, W. H., 2014. Learning and fatigue effects revisited. Investigating the effects of accounting for unobservable preference and scale heterogeneity. *Land Economics*, 90(2):323-350.

Czajkowski, M., Hanley, N., and LaRiviere, J., forthcoming. The Effects of Experience on Preference Uncertainty: Theory and Empirics for Public and Quasi-Public Environmental Goods. *American Journal of Agricultural Economics*.

Dupont, D., Price, J., and Adamowicz, W., 2014. Temporal stability of water quality values across stated preference question formats. paper presented at the Fifth World Congress of Environmental and Resource Economists, 28 June - 2 July 2014, Istanbul, Turkey.

Ferrini, S., and Scarpa, R., 2007. Designs with a priori information for nonmarket valuation with choice experiments: A Monte Carlo study. *Journal of Environmental Economics and Management*, 53(3):342-363.

Hensher, D. A., and Greene, W. H., 2011. Valuation of Travel Time Savings in WTP and Preference Space in the Presence of Taste and Scale Heterogeneity. *Journal of Transport Economics and Policy*, 45(3):505-525.

Krinsky, I., and Robb, A. L., 1986. On approximating the statistical properties of elasticities. *The Review of Economics and Statistics*, 68(4):715-719.

Liebe, U., Meyerhoff, J., and Hartje, V., 2012. Test–Retest Reliability of Choice Experiments in Environmental Valuation. *Environmental and Resource Economics*, 53(3):389-407.

Loomis, J. B., 1989. Test-Retest Reliability of the Contingent Valuation Method: A Comparison of General Population and Visitor Responses. *American Journal of Agricultural Economics*, 71(1):76-84.

MacMillan, D., Hanley, N., and Lienhoop, N., 2006. Contingent valuation: Environmental polling or preference engine? *Ecological Economics*, 60(1):299-307.

McFadden, D., 1974. Conditional Logit Analysis of Qualititative Choice Behaviour. In: *Frontiers in Econometrics*, P. Zarembka, ed., Academic Press, New York, NY, 105-142.

McNair, B., Bennett, J., and Hensher, D., 2011. A Comparison of Responses to Single and Repeated Discrete Choice Questions. *Resource and Energy Economics*, 33:554-571.

Mørkbak, M. R., and Olsen, S. B., 2014. A within-sample investigation of test–retest reliability in choice experiment surveys with real economic incentives. *Australian Journal of Agricultural and Resource Economics*:n/a-n/a.

Munro, A., and Hanley, N. D., 2002. Information, Uncertainty, and Contingent Valuation. In: *Valuing Environmental Preferences*, I. J. Bateman and K. G. Willis, eds., Oxford University Press.

Racevskis, L. A., and Lupi, F., Year. Incentive Compatibility in an Attribute-Based Referendum Model. American Agricultural Economics Association Annual Meeting, July 27-29, Orlando, Florida.

Revelt, D., and Train, K., 1998. Mixed Logit with Repeated Choices: Households' Choices of Appliance Efficiency Level. *Review of Economics and Statistics*, 80(4):647-657.

Ryan, M., Netten, A., Skåtun, D., and Smith, P., 2006. Using discrete choice experiments to estimate a preference-based measure of outcome—An application to social care for older people. *Journal of Health Economics*, 25(5):927-944.

Scarpa, R., and Rose, J. M., 2008. Design Efficiency for Non-Market Valuation with Choice Modelling: How to Measure it, What to Report and Why. *Australian Journal of Agricultural and Resource Economics*, 52(3):253-282.

Schaafsma, M., Brouwer, R., Liekens, I., and De Nocker, L., forthcoming. Temporal stability of preferences and willingness to pay for natural areas in choice experiments: A test-retest. *Resource and Energy Economics*.

Scheufele, G., and Bennett, J., 2013. Effects of Alternative Elicitation Formats in Discrete Choice Experiments. *The Australian Journal of Agricultural and Resource Economics*, 57:214-233.

Skjoldborg, U. S., Lauridsen, J., and Junker, P., 2009. Reliability of the Discrete Choice Experiment at the Input and Output Level in Patients with Rheumatoid Arthritis. *Value in Health*, 12(1):153-158.

Street, D. J., and Burgess, L., 2007. The Construction of Optimal Stated Choice Experiments: Theory and Methods. Wiley-Interscience, Hoboken, NJ.

Street, D. J., Burgess, L., and Louviere, J. J., 2005. Quick and easy choice sets: Constructing optimal and nearly optimal stated choice experiments. *International Journal of Research in Marketing*, 22(4):459–470.

Swait, J., and Adamowicz, W., 2001. Choice Environment, Market Complexity, and Consumer Behavior: A Theoretical and Empirical Approach for Incorporating Decision Complexity into Models of Consumer Choice. *Organizational Behavior and Human Decision Processes*, 86(2):141-167.

Train, K. E., and Weeks, M., 2005. Discrete Choice Models in Preference Space and Willingness-to-pay Space. In: *Applications of Simulation Methods in Environmental and Resource Economics*, R. Scarpa and A. Alberini, eds., Springer, Dordrecht, 1-16.

Vossler, C. A., Doyon, M., and Rondeau, D., 2012. Truth in Consequentiality: Theory and Field Evidence on Discrete Choice Experiments. *American Economic Journal: Microeconomics*, 4(4):145-171.

Whittington, D., Smith, V. K., Okorafor, A., Okore, A., Liu, J. L., and McPhail, A., 1992. Giving respondents time to think in contingent valuation studies: A developing country application. *Journal of Environmental Economics and Management*, 22(3):205-225.

Annex A. The comparison of attitudinal responses in surveys 1. and 2.

Table A1. The number of respondents who agreed / disagreed with the statement that "the Białowieża Forest should be protected"

Survey 2. / Survey 1.	Definitely agree	Rather agree	Neither agree not disagree	Rather disagree	Definitely disagree	Total for survey 1.
Definitely agree	632	98	3	1	1	735
Rather agree	20	28	0	1	0	49
Neither agree not disagree	2	2	0	0	0	4
Rather disagree	0	1	0	0	0	1
Definitely disagree	0	0	0	0	0	4
Total for survey 2.	654	129	3	2	1	789

(Kruskal Wallis test statistic 38.94, p-value = 0.000; Wilcoxon signed rank test p-value = 1.000)

Table A2. The number of respondents who agreed / disagreed with the statement that "the current management of National Park and Natural Reserves part of Białowieża Forest is adequate"

Survey 2. / Survey 1.	Definitely agree	Rather agree	Neither agree not disagree	Rather disagree	Definitely disagree	Total for survey 1.
Definitely agree	262	130	5	13	2	412
Rather agree	103	189	9	10	1	312
Neither agree not disagree	5	24	4	3	0	36
Rather disagree	8	13	0	2	0	23
Definitely disagree	1	2	0	0	3	6
Total for survey 2.	379	358	18	28	6	789

(Kruskal Wallis test statistic 1.28, p-value = 0.259; Wilcoxon signed rank test p-value = 0.3910)

Table A3. The number of respondents who stated that "the National Park and Natural Reserves parts of Białowieża Forest should be:"

Survey 2. / Survey 1.	Passively protected	Used commercially	Managed in some other way	Total for survey 1.
Passively protected	464	156	13	633
Used commercially	59	77	0	136
Managed in some other way	5	10	5	20
Total for survey 2.	528	243	18	789

(Kruskal Wallis test statistic 33.59, p-value = 0.000)

Table A4. The number of respondents who agreed / disagreed with the statement that "the number of visitors to the Białowieża Forest should be limited"

Survey 2. / Survey 1.	Definitely agree	Rather agree	Neither agree not disagree	Rather disagree	Definitely disagree	Total for survey 1.
Definitely agree	109	92	4	12	6	223
Rather agree	69	167	15	46	4	301
Neither agree not disagree	2	19	9	13	3	46
Rather disagree	18	63	14	62	14	171
Definitely disagree	2	7	2	26	11	48
Total for survey 2.	200	348	44	159	38	789

(Kruskal Wallis test statistic 0.01, p-value = 0.9030; ; Wilcoxon signed rank test p-value = 0.2603)

Table A5. The number of respondents who agreed / disagreed with the statement that "their household is willing to pay higher taxes to introduce changes in the protection regime of the Białowieża Forest"

Survey 2. / Survey 1.	Definitely agree	Rather agree	Neither agree not disagree	Rather disagree	Definitely disagree	Total for survey 1.
Definitely agree	17	18	3	4	2	44
Rather agree	18	139	27	51	9	244
Neither agree not disagree	1	38	36	36	7	118
Rather disagree	3	58	38	121	47	267
Definitely disagree	2	10	6	45	53	116
Total for survey 2.	41	263	110	257	118	789

(Kruskal Wallis test statistic 0.44, p-value = 0.509; ; Wilcoxon signed rank test p-value = 0.5646)

Faculty of Economic Sciences
University of Warsaw
44/50 Długa St.
00-241 Warsaw
www.wne.uw.edu.pl