Preference and WTP stability for public forest management

Abstract: The assumption of the stability of preferences is fundamental to consumer theory and the use of cost-benefit analysis. 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 one survey and between two different moments of time. The "within survey" test compares a set of responses from individuals over the sequence of the first 12 and the second 12 choices in a stated preference survey; the "between two different moments of time" test compares responses from the same people over a period of 6 months. Non-parametric analysis reveals little clear trending in choices across these sets, and higher consistency for status quo choices than for enhanced environmental management choices. Overall, we reject the strictest test of the equivalence of WTP distributions between choice sets. However, we also find that respondents' mean willingness to pay is fairly stable both within survey and between moments of time. Such differences as emerge are mainly driven by the changes in variances of WTP and by imperfect correlations of individual-specific WTP between choice sets.

JEL classification: D01, H4, Q23, Q51

Keywords: preference stability, test-retest, discrete choice experiment, contingent valuation, stated preferences, forestry

Highlights:

- Preference stability is an important assumption in welfare economics and the economic theory of value. We test this assumption using evidence from a discrete choice experiment study of forest management.
- Using the same group of people, we compare willingness to pay estimates derived from two sets of choice tasks presented in one survey, and with estimates obtained from the same people six months later.
- We analyse the dynamics of responses to choice tasks, showing how and when people changed their minds.
- We find that average willingness to pay does not significantly change both within survey and between two different moments of time. However, the shapes of the distributions may differ.

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 preferences for the same good, and willingness to pay (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 the characteristics of a good (Munro and Hanley, 2002) or one's preferences for experience goods (Czajkowski, Hanley and LaRiviere, 2014). However, in the standard model preference parameters are supposed to be stable (McFadden, 2001). This is a crucial assumption when valuation of a public good is conducted in order to inform policy makers. If preferences are unstable such that willingness to pay for a specific change in the quantity of a public or private good varies even though there is no change in any of the standard economic drivers of welfare measures, then benefit-cost analysis is no longer informative as to the efficiency implications of policy change or changes in environmental management. For example, changes in stated willingness to pay due to variations in the emotional condition of a respondent would mean that the Kaldor-Hicks potential compensation test could no longer be applied (since whether gains exceeded losses would depend on un-observable variations in context). Our study sheds some light on validity of valuation methods with regard to preference stability assumption, since we test both the stability of an individual's willingness to pay for a good across a sequence of choice tasks in an initial survey, and across a 6-month period between this initial survey and a follow-up survey.

Specific tests for preference stability over environmental goods can be found in both contingent valuation (CV) and discrete choice experiment (DCE) settings. CV test-retest procedures were conducted among others by Loomis (1989), Carson *et al.* (1997), Brouwer (2006) and Brouwer (2012). In all cases two surveys were carried out over an interval ranging from two weeks to two years. The results in all cases indicate that the average WTP is stable.

Test-retest procedures have also been applied within DCE. Bliem, Getzner and Rodiga-Laßnig (2012) estimate multinomial and mixed logit models 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, Meyerhoff and Hartje (2012) used an Error Component 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 and Olsen (2014) used DCE to compare responses over a 2 week period for a market good (apples) with "real economic incentives". They thus sought to undertake a test-retest experiment in an incentive-compatible setting. They found "very good agreement" between the DCE estimates of preferences over this rather short time interval. However, their sample consisted of 25 persons only.

Further relevant contributions include Dupont, Price and Adamowicz (2014), who compare estimates of WTP for health end points related to water quality in Canada between surveys undertaken in 2004 and 2012, using both CV and DCE. The health end points relate to illness and death cases from microbial infections and bladder cancer. They found that whilst there was a significant change in estimated 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 and Logar (2014), who survey the same sample of people in Switzerland at a 6-month interval using both CV and CE. Their study relates to WTP for upgrading of waste water treatment plants in Switzerland to remove micro-pollutants. Some 20% of CE responses and some 30% of CV responses showed 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) and Skjoldborg, Lauridsen and Junker (2009) provide test-retest analysis for preferences regarding health care.

Unfortunately, within-survey tests of preference stability within a DCE setting may also 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 the distribution of their preference type becomes narrower as experience in choosing increases (Czajkowski, Hanley and LaRiviere, 2014). As people repeat choices, they may also find that a choice task becomes simpler; or else they may become bored and start using heuristics more frequently (Swait and Adamowicz, 2001). Any of these effects could show up as a change in the estimated 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, Giergiczny and Greene, 2014). A review of multiple such "ordering effects" as well as their empirical testing can be found in Day *et al.* (2012). There have been a number of papers which also demonstrate a related "time to think" effect on WTP for changes in an environmental good (Whittington *et al.*, 1992; MacMillan, Hanley and Lienhoop, 2006).

In this paper, we conduct both within survey and between two different moments of time tests of the stability of choices and estimated distributions of WTP. These tests are based on observations of the same individuals. The within-survey test considers responses to the first 12 and then second 12 choice questions in a survey on options for forest management. The between moments of time test compares

these choices with responses from a similar (and for one subsample identical), 12-question DCE carried out six months later. This design provides a contribution to most of the test-retest literature, which as noted above has focussed on between moments of time tests only. In a within-survey experiment, 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 can confuse any signal about preference stability. This perspective stands in contrast to between moments of time tests, but here the researcher must confront a different set of problems, such as whether an individuals' socio-economic conditions changed, or where they may learn more about the good (rather than learning their preferences) over the interval. By investigating both issues jointly, our study provides an insight into the extent of the changes which may result from each of them. Although the two phenomena may be caused by different behavioural and economic effects, researchers' interest is basically the same in both cases – whether the hypothesis of stable welfare measures can be rejected (within a sequence of choices in a survey or between two different moments of time).

2. Study Design and Data

2.1. The setting of the study – the Białowieża Forest

The Białowieża Forest in Poland is an ancient woodland straddling at the border between Belarus and Poland, located in north east-central Poland on the border with Belarus. It is one of the last and largest remaining parts of the immense primeval forest which once spread across the European Plain. The Białowieża Forest is one of the most recognized and ecologically valuable forests in Poland (Czajkowski, Buszko-Briggs and Hanley, 2009). Despite some visible signs of human activity, it is still commonly considered the last natural lowland forest in temperate Europe. It is especially regarded for its natural dynamics as well as its species richness, and its ecological structures and functions (Wesołowski, 2005).

From the early 1990s biologists, environmentalists and various NGOs have been trying to convince decision makers to enlarge BNP over the entire territory of the Białowieża Forest; so far, unsuccessfully. One of the aims of conducting our study was to provide arguments in public discussions regarding the enlargement of the Białowieża National Park and possible changes in the forest management. In addition, our survey was constructed in a way which enabled testing preference stability, which is the main purpose of this paper.

A few one-to-one in-depth interviews were conducted by the research team members to fine-tune the survey instruments (structure, wording, visual materials – maps and photos). After consultations with

biologists[1] working in the Białowieża Forests two possible management levels for the forests outside BNP and the reserve have been considered that is:

1) maintain the current management typical for managed forest or
2) enlarge the passive protection zone, that is to allow for rewilding[2] of the managed part of the Białowieża Forest.

It was explained that these options would result in low or high level of forest naturalness respectively.

The differences between managed forests (low level of naturalness) and close to natural forests (high level of naturalness) were explained to the respondents with the use of photographs, drawings and written descriptions presented in Figure 1.

The Białowieża Forest can be divided into three relatively homogenous parts which differ in naturalness levels. The short characteristic and possible changes in the management in each part were explained to the respondents.

1) **The Białowieża National Park and the nature reserve Natural Forests of the Białowieża Primeval Forest** – the core of the Białowieża Forests is 10,500 ha protected in the Białowieża National Park (16% of the Polish part of the forest). In addition, 12,000 ha (19% of the Polish part of the forest) of natural forests are protected in the reserve outside the Białowieża National Park. Almost the entire area of the Białowieża National Park and forests in the reserve are protected forests under non-intervention regime. The Białowieża National Park and the reserve comprise the best preserved part of the Białowieża Forest. Some species of lichens, fungi, insects which depend on high volumes of dead wood are present only here (Wesołowski, 2005). In the questionnaire we referred to this part as having 'High naturalness level'.
2) **Second growth forests** – approximately 6,000 ha (15% of the Polish part of the Białowieża Forest) are so called remains after large scale clear-cuts made in 1920's by the British European Century Timber Corporation known as Centura which made no attempt to renew these clear-cuts[3] (Directorate General of the State Forests, 2011). These plots have been naturally regenerated and in most of these places no significant human intervention have taken place over the last 90 years. However, this may change at any time as these stands are considered by the State Forests as commercial forests. It was explained in the questionnaire that the second growth

[1] We are very grateful to prof. Bogdan Jaroszewicz a director of Białowieża Geobotanical Station for his comments on an early draft of our questionnaire.

[2] By rewilding we mean the whole process of returning ecosystems to a state of ecological health and dynamic balance, making them self-sustaining, without the need for ongoing human management (Navarro and Pereira, 2012).

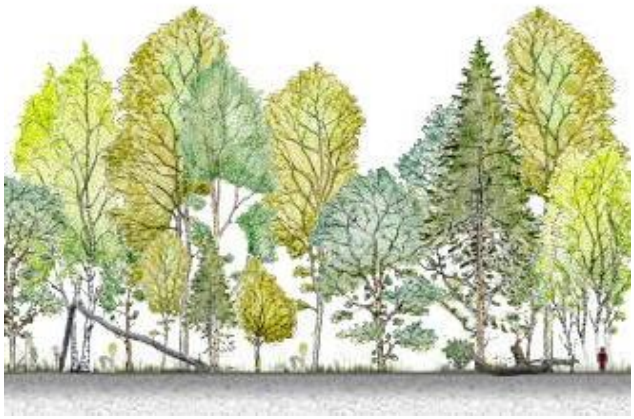
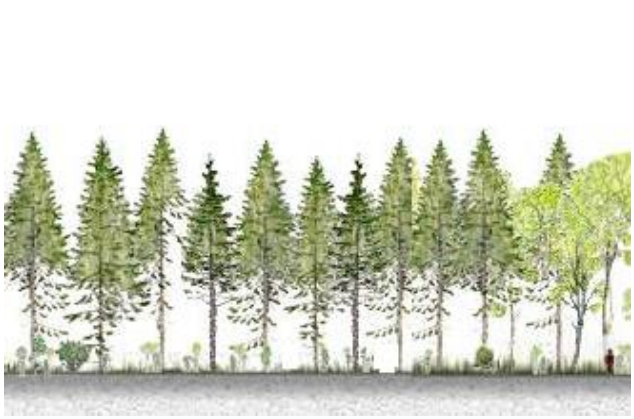
[3] Centura was obliged to use 10-hectare clear-cuts and retain seed trees. In fact, the clear-cuts often exceeded a hundred hectares each and no attempt was made at their renewal (Directorate General of the State Forests, 2011).

forests are under dynamic changes now as the climax status has not been reached yet and that their future status will depend on the chosen management. If passive protection is extended this forest will have 'High naturalness level' as the forests in the Białowieża National Park and the reserve outside the Park currently. After consultations with forest botanists we informed the respondents that this status will be reached in about 150 years. However, if these forests are turned into standard managed forests they will have 'Low naturalness level' as the managed forests in the Białowieża Forest.

Managed forests – approximately 33,000 ha (50% of the Polish part of the Białowieża Forest) are standard managed forests. In the questionnaire we referred to this part as having 'Low naturalness level'. After consultations with forest botanists we informed the respondents that it is possible to rewild these forests that is if passive protection is extended over the managed forests they will have 'High naturalness level' as the forests in the Białowieża National Park and the reserve outside the Park currently. The respondents that this status will be reached in about 250 years

Finally, 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 been 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 in the summer the forest is visited much more frequently, with up to 10,000 people on exceptionally busy days. In the future, these numbers may increase. On the one hand, a large number of visitors may make the achievement of a higher level of forest naturalness difficult, and may even reduce the current level of ecological quality. For many years, environmentalists and various non-government organizations 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.

Figure 1. The differences between natural forests and managed forests

Natural forest (high level of naturalness)	**Managed forest (low level of naturalness)**
	
	
1) All trees are left in the forest until they die and decay. 2) Trees regenerate naturally by self-sown seeds 3) Forest has a multi-age structure 4) There are usually many tree species 5) High volume of dead wood (100 m3/ha or more). 6) There is a greater diversity of species of plants, animals and fungi. Many rare species can live only in the forests with a large quantity of old rotting trees.	1) After attaining a certain age, the forest is logged. Old trees a met rarely. 2) Trees regenerate artificially by direct seeding or *planting*. 3) Forest has an even-age structure 4) Usually scots pine or spruce monoculture 5) There is a small volume of dead wood (less than 10 m3/ha). 6) There is a much smaller diversity of species of plants, animals and fungi. Rare species do not have good conditions to live here.
Close to natural forests cover about 0.6% of all Polish forests	About 99% of forests in Poland are managed forests

2.2. Experimental design and data

The DCE comprised five attributes, of which one – forest naturalness in the National Park and the nature reserve – could take only one level and was presented for completeness. 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, and thus the level of this attribute was held constant across all choice tasks and alternatives. The next two attributes referred to future management programs (and the resulting naturalness level of that part of the Białowieża Forest) for the areas which are currently under commercial management, and the second-growth sections of the Białowieża Forest. These management programs (attribute levels) were "commercial use" or "passive protection", resulting in a low and high level of naturalness in the future, respectively. The forth attribute was the maximum visitor numbers allowed for the entire Białowieża Forest, i.e. limiting the tourism pressure on the forest. Lastly, the design included the cost attribute. The payment vehicle was described as the increase in income taxes paid annually by each household in Poland. The attribute and attribute levels are summarized in Table 1.

Table 1. Attributes and levels used in the DCE study

Attributes	Levels	Variables
National Park and Nature Reserve (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	VIS_1
	7,500 visitors per day	VIS_2
Annual cost per household	0, 25, 50, 75, 100 PLN	*COST*

The contingent scenario description, attributes and their levels were developed in the process of extensive qualitative testing to make sure the survey is understandable and the program credible. The main survey was preceded with a pilot study, aimed at making sure the instrument worked well and for collecting priors for the main study design.

The data used in this study was collected through a national online survey of the Polish population using two surveys carried out by a professional polling agency. The first survey was conducted in December 2011. Of the 3,016 respondents who were invited to take part in the web panel survey 1,459 responded,

and 1,302 of them provided usable responses. The final sample was quota-controlled for sex, age, region and city size. In June 2012, 998 of the participants who successfully completed the first survey were invited to participate in a follow-up survey, and 789 of them successfully completed the survey. Since we were able to identify specific respondents within the data set, this provides an opportunity for us to test how their preferences changed over the course of the 6-month period (December to June) while controlling for the panel structure of the data.[4]

The first survey (December 2011) included 24 choice tasks, which were in fact the same 12 choice tasks repeated. The first 12 choices are hereafter referred to as set A and the second 12 choices as set B. This design provides an opportunity for us to undertake a within-sample test of preference stability, and indeed to compare on a choice-task by choice-task basis. The second survey (June 2012) included 12 choice tasks (set C). The design of the study is illustrated in Figure 2. Comparing set A with set C for each person is thus a "between two different moments of time" comparison of preference stability.[5]

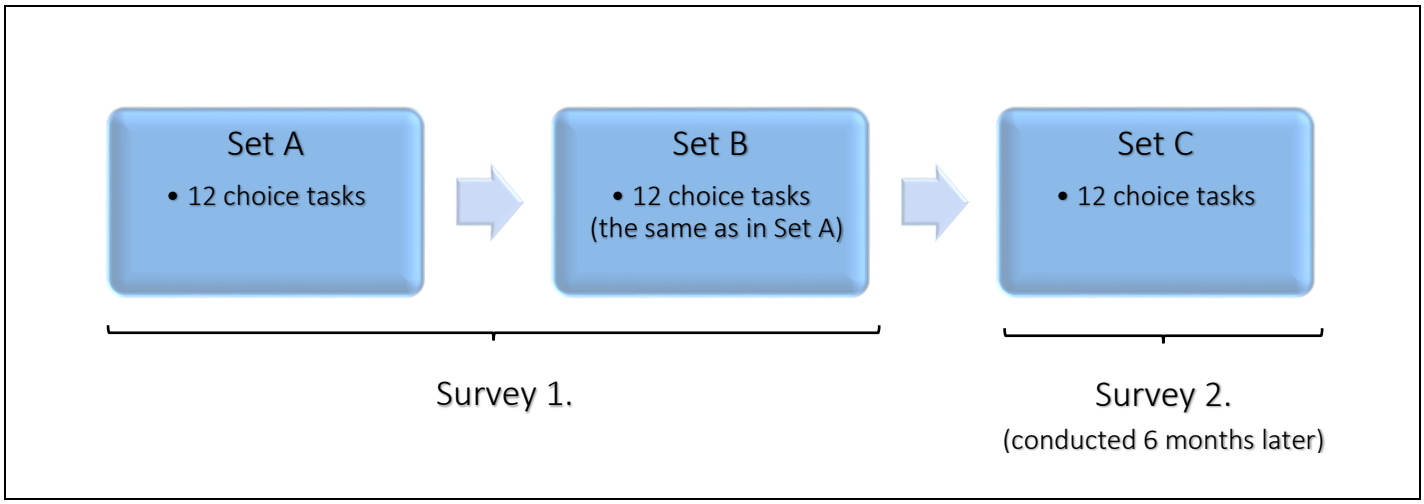


Figure 2. Illustration of the design of the preference stability study

The experimental design of our study utilized both an optimal-in-difference (Street, Burgess and Louviere, 2005; Street and Burgess, 2007) and an efficient design (Ferrini and Scarpa, 2007; Scarpa and Rose, 2008) approach.[6] The designs were generated and applied separately for half of the participants

[4] The questionnaire and dataset used in this study are made available online at czaj.org. The models were estimated using custom code developed in Matlab which is made available from github.com/czaj/DCE under Creative Commons BY 4.0 license.

[5] Even though the same choice tasks were repeated in the same survey, respondents of the pretesting phase and the main survey appeared unaware of this, or they were not disturbed and did not comment on this.

[6] 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

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. For ¼ of our respondents (n = 193) we are thus additionally able to test if their *choices* changed between the first and the second survey, in addition to testing if their *implied preferences* changed.

Each choice task consisted of 3 alternatives, one of which was the Status Quo (*SQ*, meaning no changes in current management program). 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. We have taken steps to ensure that each choice task and each alternative was presented in every position in the sequence a comparable number of times. An example of a choice card is presented in Figure 3.

	Program A	**Program B**	**Program C**
	Continuation of current management program	Changes in current management program	Changes in current management program
National Park and Natural Reserves (35% of the Białowieża forest)	High level of naturalness	High level of naturalness	High level of naturalness
Commercial forests (50% of the Białowieża forest)	Low level of naturalness	Low level of naturalness	High level of naturalness
Second-growth forests (15% of the Białowieża forest)	Low level of naturalness	High level of naturalness	High level of naturalness
Number of visitors (per day)	No limit	No limit	5,000
Cost for your household (per year)	0 PLN	50 PLN	100 PLN
Your preferred program:	☐	☐	☐

Figure 3. An example of a choice card presented to the respondents.

3. Non-parametric analysis

In what follows, 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

employed Bayesian approach (Bliemer, Rose and Hess, 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.

if choices differed in any pair of sets. Additionally we checked whether a pattern of choosing the *SQ* alternative differ across sets.

3.1. Within survey comparison

Figure 4 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.[7]

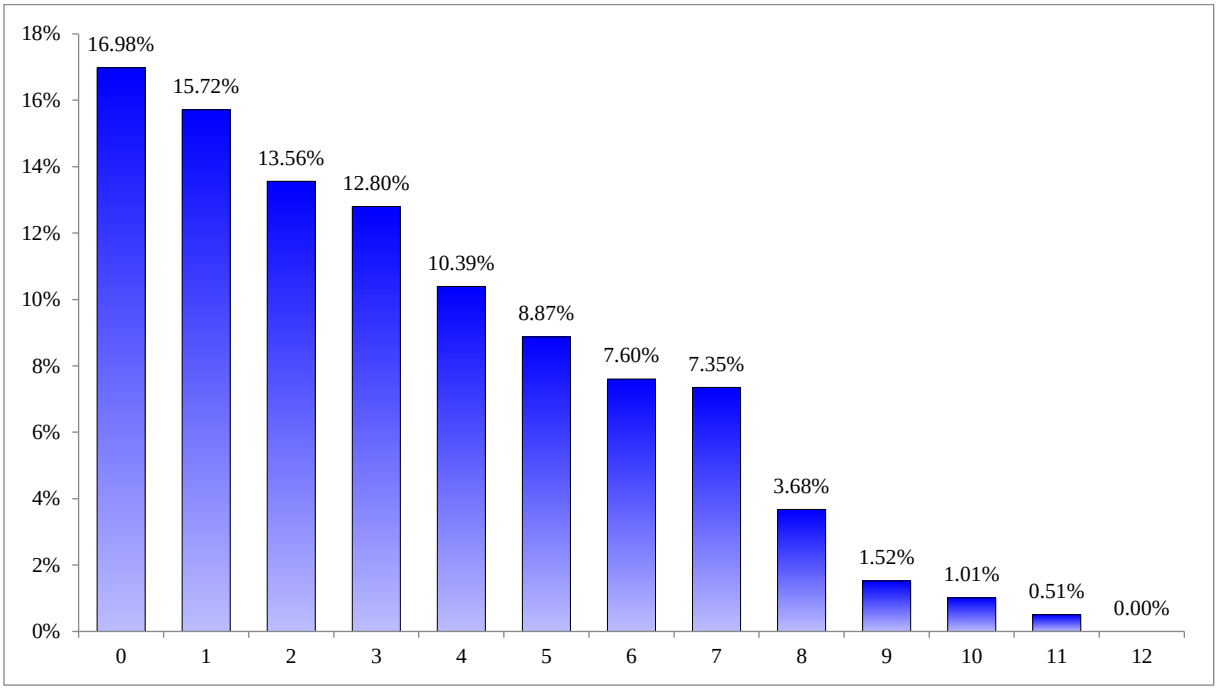


Figure 4. The distribution of respondents who changed their decisions a particular number of times in the series of 12 choice tasks repeated twice (sets A and B)

Additional insights are provided by an analysis of the dynamics of decision changes illustrated in Figure 4. In addition to the share of unchanged decisions in each choice task, we present their 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 when combined with the relatively high degree of variability in respondents' choices (a low scale

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

parameter, i.e. a high utility function error term variance in relation to its deterministic component). 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 in the sense that both groups seemed equally likely to change their decisions at every choice task. 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 5.

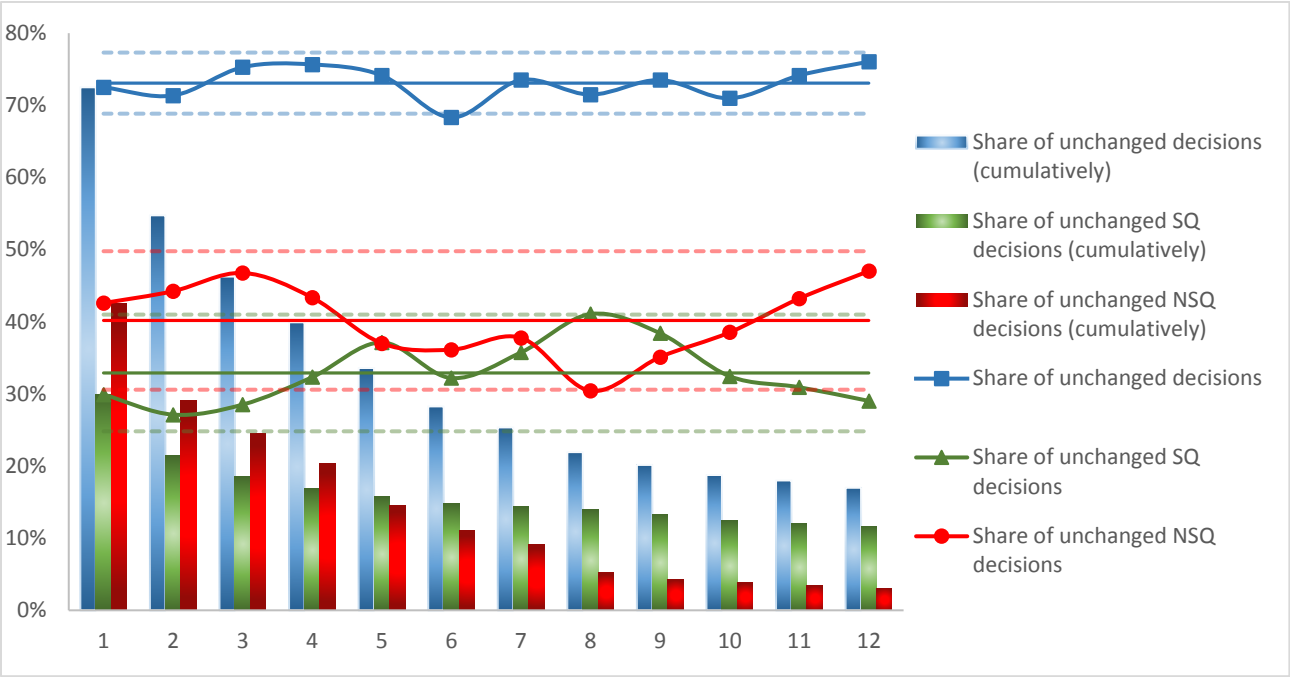


Figure 5. Dynamics of the number of decision changes in different choice tasks (set A vs. B; n = 789)

In Annex A we additionally included analogous plot for decision changes dynamics between sets A and B for respondents who did not participate in second survey. The plots are very similar, there seems to be no significant differences between them. To formally confirm this, we applied the Kruskal-Wallis test of the differences in shares of unchanged decisions (overall, *SQ* and *NSQ*) for each choice task. The results show no apparent patterns, the only significant differences were observed for the fifth choice task for *SQ* shares and the eighth choice task for the *NSQ* shares. We therefore conclude that there are no major differences between these samples with regard to preference dynamics and there is no apparent self-selection bias, which could influence our results.

3.2. Between two different moments of time differences

Figures 6 and 7 present the dynamics of decision changes between surveys (i.e., between sets A and C, and between sets B and C, respectively). Overall, there is more variation in these cases, 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. Like in the within survey case, the dynamics of respondents' choices do not reveal any significant patterns. Interestingly, the only respondents who answered consistently between the sets A and C were those who always chose the *SQ* alternative.

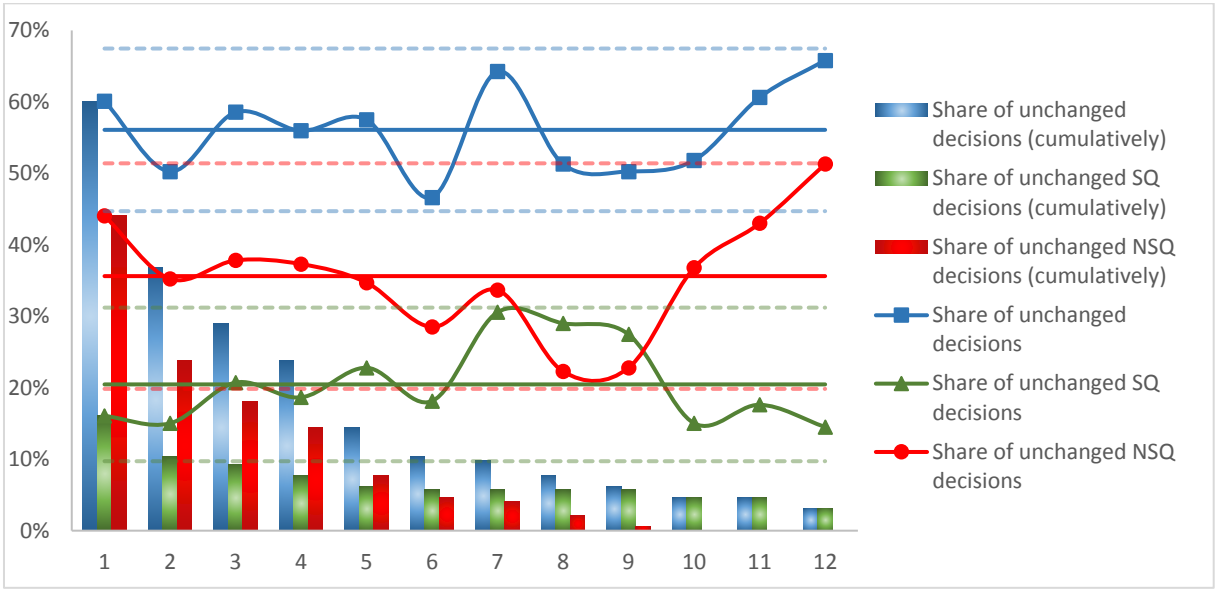


Figure 6. Dynamics of the number of decision changes in different choice tasks (set A vs. C; *n* = 193)

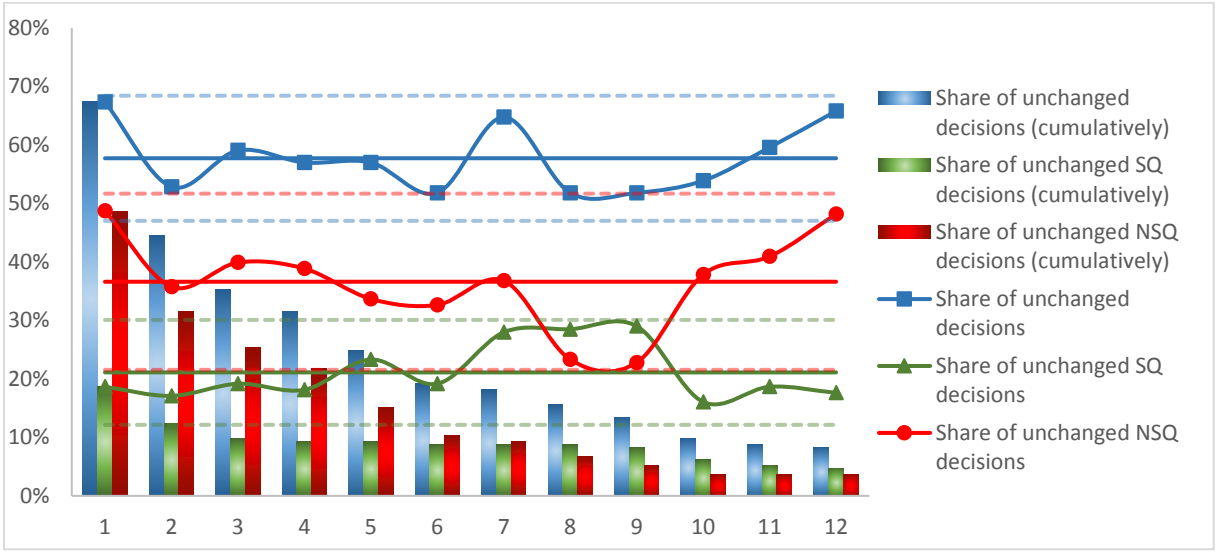


Figure 7. Dynamics of the number of decision changes in different choice tasks (set B vs. C; *n* = 193)

Finally, we analyze the share of the SQ alternative choices. The respondents who had selected the SQ alternative in every choice task constitute 14.70%, 19.52% and 15.46% of the sample for sets A, B and C, respectively. The SQ alternative could have been selected by some respondents in some choice tasks only. Figure 8 presents the shares of the SQ alternative choices in each choice task. As argued by Day *et al.* (2012), the occurrence of fatigue can be characterized by, among others, an increase in the probability of selecting the *SQ* alternative. In our case, the share of respondents choosing the *SQ* alternative in each choice task is very similar. Despite some increase between the third and the eighth choice task, and subsequent decrease in the probability, there is no significant difference in the share of *SQ* choice between set A and B. We interpret it as an indication of no apparent fatigue effect of this type, although it should be noted that in set B share of respondents who have chosen only SQ increased when compared with set A.

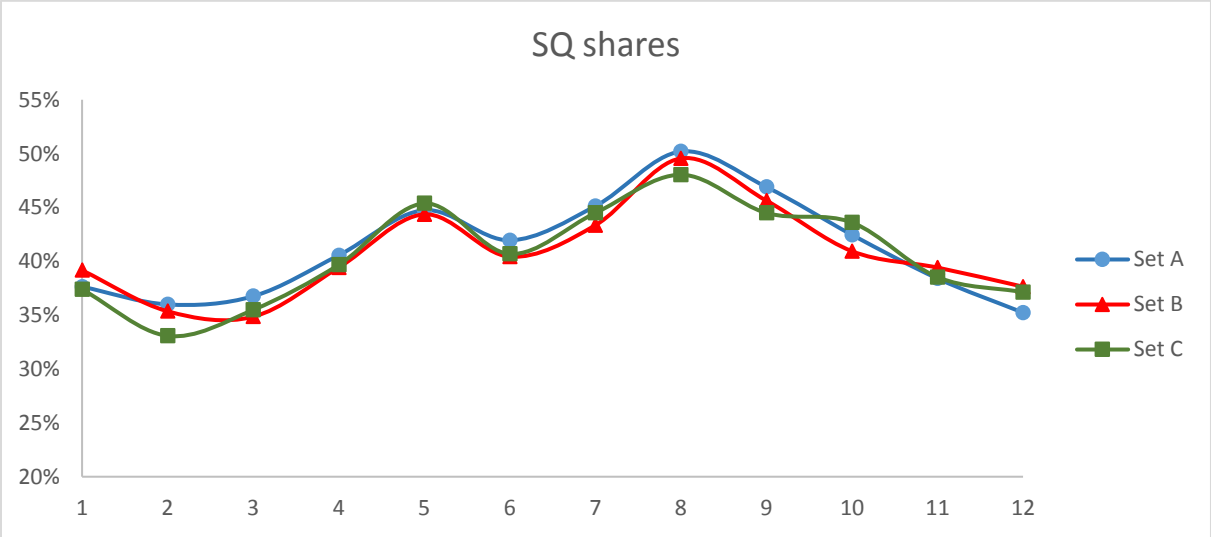


Figure 8. Shares of respondents who've chosen status quo alternative. (sets A, B and C; n = 789)

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' marginal WTP distributions were stable between groups of choice sets A and B presented in survey 1, and between these choice sets and choice

set C presented in survey 2 six months later. We focus on stability of welfare measures instead of stability of preferences, as it is usually done, for two reasons. First, WTP is the main result of valuation studies, the reason most of these studies are conducted, and in the focus of policymakers who wish to undertake a cost-benefit analysis. Second, WTP measures are scale-free and therefore can be directly compared. For the comparison of preference parameters Swait and Louviere (1993) or Czajkowski, Hanley and LaRiviere (2016) approach should be applied to take the possible differences in scale between different sets into account. Having 3 sets and using this procedure for mixed logit is computationally cumbersome.[8]

To conduct our analysis we used the mixed logit model (MXL, Revelt and Train, 1998) which was estimated in the so called WTP-space (Train and Weeks, 2005), which means that utility of *n*-th individual from choosing *i*-th alternative in *j*-th choice task is reparametrized in following way:

$$U_{ijn} = \mathbf{X}_{ijn}\mathbf{b}_n - \alpha_n c_{ijn} + \varepsilon_{ijn} = \alpha_n\left(\mathbf{X}_{ijn}\,\mathbf{b}_n/\alpha_n - c_{ijn}\right) + \varepsilon_{ijn} = \alpha_n\left(\mathbf{X}_{ijn}\boldsymbol{\beta}_n - c_{ijn}\right) + \varepsilon_{ijn}$$

where, c_{ijn} denotes cost of given alternative, $\mathbf{X}_{ijn}$ is a vector of other attributes and ε_{ijn} is a random term. In this framework, the parameters for all attributes ($\boldsymbol{\beta}_n$) but the cost can directly be interpreted as WTP (in PLN per year)[9]. The advantage of this approach is that no additional calculations are needed to estimate some characteristics of WTP distribution, e.g., means of normally distributed random parameters can simply be interpreted as mean WTP. In this model we assumed that all parameters are random, assuming that WTP for all attributes follow normal distribution, while the negative cost parameter (α_n) is log-normally distributed.[10] 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).

In order to analyse stability of WTP we interacted every attribute with binary variables for sets A, B and C, which lead to 18 random parameters in the model. Evaluating the statistical significance of the differences of the estimates for different sets provides a convenient way of testing if respondents' preferences as revealed in sets A, B and C are stable. In addition, we also allowed for full correlation between all random parameters. This is important for two reasons: (i) it is very likely that WTP for the same attributes in different sets are highly correlated (in the case of no changes in respondents' choices these correlations should be equal to one), and (ii) the (possibly) heteroskedastic variance of error terms

[8] Particularly, because a meaningful likelihood ratio test comparisons of models estimated with the simulated maximum likelihood method requires using very many draws (Andersen, 2014).

[9] 1 PLN ≈ 0.23 EUR ≈ 0.25 USD

[10] We also tested the specification in which WTP followed log-normal distributions. It was not better fitted to the data, however (Voung test statistic = -0.4028).

could be different between sets A, B and C, but it is also very likely that its distribution is correlated between them.[11]

Table 2 presents the results of the MXL model. Estimation was performed using Matlab 8.1. The maximum likelihood function was simulated using 10,000 scrambled Sobol draws (Czajkowski and Budziński, 2015). Standard errors of coefficients associated with standard deviations of random parameters were simulated using Krinsky and Robb method with 10^6 draws (Krinsky and Robb, 1986).

The results presented in Table 2 show that, overall, respondents were generally in favor of extending passive protection to commercial (*COM*) and second-growth (*SGR*) areas of the Białowieża Forest, with average WTP in the range of 12.65 to 17.12 PLN for the former and 21.77 to 29.99 PLN for the latter. The distributions of these WTP were highly heterogeneous, as indicated by significant standard deviations of the random parameters. We can also observe a dislike regarding the *SQ* alternative (negative but heterogeneous WTP associated with the alternative specific constant for the no-change alternative). WTP for restricting the number of visitors (VIS_1, VIS_2) were highly heterogeneous as well, but mostly insignificant with different signs for different sets. As for the dynamics, we can observe that mean WTP for the *SQ*, *COM* and *SGR* is higher in set B than in set A, but then decreases to an even lower level in set C. WTP associated with VIS_1 and VIS_2 presents the opposite pattern, but due to the insignificance of these parameters it is not very informative. We revisit these conclusions later, in more formal way, using Wald test.

Inspecting the correlation matrix of random parameters (which are equivalent to WTP for all attributes except cost), estimated with the MXL model, presented in Table 3, provides additional insights. The greyed-out cells correspond to correlations between parameters of the same attributes in different sets. The correlations between WTP in set A and their counterparts in set B are very high, ranging from 92.3% to 99.6%. This means that WTP in set B are approximately linear transformations of their set A equivalents.[12] The correlations between sets A and C as well as sets B and C are much lower, but still positive and significantly different from zero.[13] WTPs are highly correlated within every set as well, especially WTP for VIS_1 and VIS_2, which was expected.

[11] For example, respondents who were observed to have had a high variance of the error term in set A are also likely to have had a high variance in set C.

[12] Note that it is not a "scale effect", as WTP's are scale free.

[13] The correlations between the random *COST* parameters is lower than what we observe for WTP. However, because these parameters are a product of scale (set specific) and the parameter representing cost sensitivity, and it is not possible to disentangle the two effects, this observation offers limited insight and is difficult to interpret – we cannot identify the driver of the observed changes in heterogeneity structure.

Table 2. The results of the mixed logit model accounting for preference and scale differences between sets A, B and C (in WTP-space).

	Mean WTP (s.e.)			**Standard deviation of WTP (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)**
SQ	-33.5013***	-31.0284***	-38.2758***	93.4651***	95.8237***	99.2762***
	(1.7048)	(1.9827)	(2.4345)	(4.9807)	(4.8928)	(5.5029)
COM	14.4530***	17.1193***	12.6514***	30.5130***	45.2908***	38.1746***
	(1.3746)	(1.7130)	(1.6731)	(1.4613)	(2.2754)	(2.0322)
SGR	28.5644***	29.9556***	21.7744***	35.5747***	49.9512***	36.6213***
	(1.4058)	(1.8671)	(1.7015)	(1.5599)	(2.1923)	(1.6475)
***VIS*1**	0.7744	-2.9075	0.4340	19.9192***	25.5335***	19.3573***
	(1.5689)	(1.9450)	(1.9776)	(2.0502)	(2.0286)	(1.6864)
***VIS*2**	-3.0014*	-2.5435	-2.6190	41.1142***	47.2967***	40.4539***
	(1.7697)	(2.1626)	(2.4219)	(2.1927)	(2.0489)	(2.1286)
COST	-2.0903***	-1.7978***	-1.9872***	1.1387***	1.4160***	1.1986***
(pref. space equivalent)	(0.0470)	(0.0595)	(0.0520)	(0.0474)	(0.0693)	(0.0647)
Log-likelihood (constants only)		-30,833.5642				
Log-likelihood		-18,496.7692				
McFadden R^2		0.4001				
Ben-Akiva R^2		0.5449				
AIC/*n*		1.3158				
n (observations)		28,404				
k (parameters)		189				

Note: ***, **, * represent statistical significance at the 1%, 5% and 10% levels, respectively.

Table 4 presents the results of the Wald tests conducted for different hypotheses of parameter equality. It consists of 4 panels. In the first three, we tested whether WTP means, variances and means and variances jointly are equal between every pair of sets, and all sets jointly for each of attributes, and additionally – if they are equal for all the attributes jointly. The fourth panel provides the results for joint equality of WTP means and variances and their respective correlations between sets being equal to 1.

Formally, we could say that the distributions of WTP are stable if we were not able reject the null hypothesis of the equality of random parameters means and variances for each of attribute between all sets (A = B = C) and if the correlations of random parameters were perfectly correlated between the 3 sets. The Wald test statistic for such a hypothesis is 582.46 which lead to its rejection (d.f. = 30). A less restrictive definition of stability involves the equality of means and variances of WTP.[14] In this case, the Wald statistic is 80.30 which also indicates the null hypothesis of equality can be rejected (d.f. = 20). In conclusion, we cannot say that the distributions of welfare measures are stable between all sets.

[14] This hypothesis involves the equality of population-level characteristics of the WTP distributions, but does not require individual-specific WTP to be equal between sets.

Table 3. Correlation matrix of random parameters implied by the MXL model

		Set A (Survey 1)						**Set B (Survey 1)**						**Set C (Survey 2)**					
		SQ	***COM***	***SGR***	***VIS*$_1$**	***VIS*$_2$**	***COST***	***SQ***	***COM***	***SGR***	***VIS*$_1$**	***VIS*$_2$**	***COST***	***SQ***	***COM***	***SGR***	***VIS*$_1$**	***VIS*$_2$**	***COST***
Set A (Survey 1)	***SQ***	**1.0000**	**-0.5601**	**-0.5288**	**-0.8579**	**-0.8507**	0.1352	**0.9963**	**-0.5742**	**-0.6219**	**-0.8346**	**-0.8732**	**0.1091**	**0.6463**	**-0.3478**	**-0.3181**	**-0.4337**	**-0.4662**	0.0380
	COM		**1.0000**	**0.4898**	0.1080	**0.1874**	-0.0150	**-0.5613**	**0.9804**	**0.5620**	0.1454	**0.2027**	-0.0201	**-0.3048**	**0.3569**	**0.2738**	0.0492	**0.0950**	**0.2212**
	SGR			**1.0000**	**0.4734**	**0.4756**	**-0.1366**	**-0.5108**	**0.4883**	**0.9798**	**0.4223**	**0.5308**	0.0044	**-0.3195**	**0.3028**	**0.3894**	0.1329	**0.1092**	0.0535
	***VIS*$_1$**				**1.0000**	**0.9075**	**-0.2050**	**-0.8441**	0.1360	**0.5252**	**0.9231**	**0.9434**	**-0.1722**	**-0.5584**	**0.2366**	**0.2693**	**0.4328**	**0.4504**	**-0.1456**
	***VIS*$_2$**					**1.0000**	**-0.0979**	**-0.8255**	**0.2197**	**0.5549**	**0.9237**	**0.9647**	0.0087	**-0.5755**	**0.2636**	**0.2365**	**0.5360**	**0.5920**	-0.0372
	COST						**1.0000**	**0.1191**	0.0636	-0.0534	-0.0723	**-0.1833**	**0.7854**	**0.0966**	-0.0285	**0.2900**	0.1109	-0.0448	**0.4147**
Set B (Survey 1)	***SQ***							**1.0000**	**-0.5748**	**-0.6029**	**-0.8119**	**-0.8475**	**0.1009**	**0.6513**	**-0.3328**	**-0.3162**	**-0.4336**	**-0.4539**	0.0593
	COM								**1.0000**	**0.5792**	**0.1844**	**0.2164**	0.0180	**-0.2892**	**0.3380**	**0.2945**	0.0680	**0.0944**	**0.1928**
	SGR									**1.0000**	**0.4970**	**0.5926**	0.0697	**-0.3688**	**0.3316**	**0.4221**	**0.1982**	**0.1705**	0.0782
	***VIS*$_1$**										**1.0000**	**0.9568**	-0.0220	**-0.5005**	**0.1442**	**0.2554**	**0.4422**	**0.4506**	0.0299
	***VIS*$_2$**											**1.0000**	-0.0680	**-0.5654**	**0.2556**	**0.2470**	**0.4604**	**0.5024**	-0.0043
	COST												**1.0000**	0.0074	0.0132	**0.2125**	0.0857	-0.0298	**0.3722**
Set C (Survey 2)	***SQ***													**1.0000**	**-0.5714**	**-0.4637**	**-0.5963**	**-0.7142**	**0.1380**
	COM														**1.0000**	**0.3651**	-0.0601	**0.1903**	-0.0569
	SGR															**1.0000**	**0.3075**	**0.4493**	**0.1567**
	***VIS*$_1$**																**1.0000**	**0.8506**	0.0075
	***VIS*$_2$**																	**1.0000**	-0.0583
	COST																		**1.0000**

Note: Coefficients in bold are significant at at least 5% level.

Table 4. Wald test statistics (degrees of freedom in brackets) for equality of different sets of parameters in the MXL model

	Means (Panel I)				Variances (Panel II)			
	A = B	**A = C**	**B = C**	**A = B = C**	**A = B**	**A = C**	**B = C**	**A = B = C**
SQ	1.0138	2.7056*	5.6417**	5.6518*	0.1118	0.5788	0.2236	0.5788
	(1)	(1)	(1)	(2)	(1)	(1)	(1)	(2)
COM	2.1128	0.7258	3.7172*	4.0337	26.2436***	8.3517***	5.2064**	32.1968***
	(1)	(1)	(1)	(2)	(1)	(1)	(1)	(2)
SGR	0.4459	9.7694***	10.6682***	12.7026***	29.3349***	0.1169	22.3449***	30.0529***
	(1)	(1)	(1)	(2)	(1)	(1)	(1)	(2)
***VIS*$_1$**	2.3144	0.0186	1.5134	2.5617	3.5469*	0.2496	5.9240**	6.0582**
	(1)	(1)	(1)	(2)	(1)	(1)	(1)	(2)
***VIS*$_2$**	0.0313	0.0176	0.0006	0.0365	4.5015**	0.1045	5.6861**	6.7840**
	(1)	(1)	(1)	(2)	(1)	(1)	(1)	(2)
All WTP	6.9604	10.4590*	18.0242***	22.1515**	37.9044***	11.3648**	26.9242***	52.5862***
	(5)	(5)	(5)	(10)	(5)	(5)	(5)	(10)
	Means & Variances (Panel III)				**Means & Variances & Correlations (Panel IV)**			
	A = B	**A = C**	**B = C**	**A = B = C**	**A = B**	**A = C**	**B = C**	**A = B = C**
SQ	1.0173	3.5412	5.9852*	6.2237	5.0550	136.6064***	135.3316***	171.0386***
	(2)	(2)	(2)	(4)	(3)	(3)	(3)	(6)
COM	29.8657***	8.4317**	9.6575***	37.2137***	32.3881***	141.8678***	150.9603***	161.9812***
	(2)	(2)	(2)	(4)	(3)	(3)	(3)	(6)
SGR	29.8734***	9.8580***	35.0110***	42.7886***	31.5231***	117.3079***	118.7153***	129.6190***
	(2)	(2)	(2)	(4)	(3)	(3)	(3)	(6)
***VIS*$_1$**	4.2198	0.3403	6.0724**	6.7376	10.0569**	20.7889***	29.7390***	34.6797***
	(2)	(2)	(2)	(4)	(3)	(3)	(3)	(6)
***VIS*$_2$**	6.8311**	0.1048	6.9301**	9.1927*	11.9285***	75.1860***	94.1143***	102.3977***
	(2)	(2)	(2)	(4)	(3)	(3)	(3)	(6)
All WTP	43.9139***	30.1523***	43.4875***	80.2975***	57.1342***	463.9029***	482.8585***	582.4573***
	(10)	(10)	(10)	(20)	(15)	(15)	(15)	(30)

Note: ***, **, * represent statistical significance at the 1%, 5% and 10% levels, respectively.

Further investigation of the results presented in Table 4 reveals some interesting facts. First of all, WTP for the *SQ* (alternative specific constant) seems to be the most stable between all sets. As can be seen in the third panel, we cannot reject the null hypothesis of equality of mean and variance between all 3 sets (Wald test statistic = 6.22; d.f. = 4). In addition, the fourth panel reveals that the correlation of WTP for the *SQ* between sets A and B is also not statistically different from 1. This does not hold between sets A and C or B and C – the correlations are significantly lower. This result is in line with what we observed in section 3, i.e., that *SQ* choices appeared more stable (as indicated by cumulative plots and the shares of the *SQ* alternative choices). It is contrast with the result of Dekker, Koster and Brouwer (2014) who found that WTP for the *SQ* is likely to be decreasing across choice tasks.

The Wald tests results indicate that the changes in mean WTP between sets A and B are not statistically significant. Nevertheless, there are significant differences between the variances of random parameters (see Panel II of Table 4). As illustrated by Table 2, standard deviations of WTP are higher in set B than in set A. As a result, although the means of WTP do not seem to change, their standard deviations increase. In addition, we note very high correlations between the observed WTP (above 92%). This indicates that even though the distributions of WTP are getting more disperse in set B than in set A, individual respondents' WTP generally stay close.

Mean WTP do not significantly change between sets A and C either, except for the second-growth areas (*SGR*). The hypothesis of equal variances cannot be rejected on 1% confidence level in this case, due to significant difference of variances of WTP for commercial forests (*COM*). The differences between sets B and C are even more evident – means and variances are significantly different irrespectively of whether they are tested separately or jointly. Overall, we cannot conclude that WTP remain stable between sets A, B and C, although the differences seem to be mainly driven by the changes in variances and imperfect correlations of individual-specific WTP.

5. Conclusions

The temporal stability of preferences is a key working hypothesis in cost-benefit analysis 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 WTP. Between two different moments of time tests examine such changes.

But we can also look at how stated preferences and WTP evolve within an experiment if a DCE approach is taken, where people complete multiple choice tasks. Here, testing for preference stability is complicated by a different set of considerations. Apparent 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. Unfortunately, the data collected in the present survey do not allow us to isolate any learning or fatigue effects in the within-sample tests; whilst there may be changes in respondents'

circumstances and knowledge over the 6 months which elapsed between surveys 1 and 2 which we cannot control for.[15]

Our contribution is to combine within- and between two different moments of time tests of preference stability for the same set of individuals. The within-survey test compares WTP between the first 12 choices and the second set of 12 choices within the same survey. The between two different moments of time 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. Results show that respondents often changed 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 SQ.

The comparison of mean WTP revealed that this key variable is relatively stable, both within- and between two different moments of time. However, we found that the variances of WTP distributions did differ significantly, especially within survey. In addition, the analysis of correlations of individual-specific WTP showed that they were not perfectly correlated, although the correlation coefficients within survey were very high (above 92%) and between two different moments of time (observed from the same respondents 6 months later) remained positive (above 35%) and significant, indicating a high level of self-consistency. Overall, our findings led us to rejecting the hypothesis of perfect preference stability, which would here imply failing to reject the equivalence of the WTP distributions in choice sets A, B and C. This may be seen as being in contrast with the results of some earlier studies. However, this null hypothesis is more stringent than that usually applied (e.g., a simple comparison of mean WTP). We also take preference heterogeneity and correlations into account.

As a more general conclusion, despite observing statistically significant differences in WTP distributions and imperfect correlation of individual-specific WTP, our results indicate that the extent of changes is relatively low. That is, when compared with the level of uncertainty associated with welfare measures which results from sampling, survey administration, model specification or the valuation method used, the instability in willingness to pay values which are measured in our survey seem rather small. This is probably the most relevant aspect of our work for policy and management applications, and can be considered reassuring, since it implies that mean WTP remained relatively stable, both within and between surveys. This suggests some support for the continued use of cost-benefit analysis to help inform environmental management and policy decisions.

[15] Another potential source of differences was the time of the year when the two surveys were conducted. While the first one was administered in winter, the second elicited respondents' preferences 6 months later – in the summer. We acknowledge that this could also potentially influence the relative importance respondents attach to forests, and hence their WTP. This is a caveat future studies should avoid.

References

Andersen, L. M., 2014. Obtaining Reliable Likelihood Ratio Tests from Simulated Likelihood Functions. *PLoS ONE*, 9(10):e106136.

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, 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., and Budziński, W. (2015). "An insight into the numerical simulation bias – a comparison of efficiency and performance of different types of quasi Monte Carlo simulation methods under a wide range of experimental conditions." In: *Environmental Choice Modelling Conference*, Copenhagen.

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., 2014. The Effects of Experience on Preferences: Theory and Empirics for Environmental Public Goods. *American Journal of Agricultural Economics*, 97(1):333-351.

Czajkowski, M., Hanley, N., and LaRiviere, J., 2016. Controlling for the effects of information in a public goods discrete choice model. *Environmental and Resource Economics*, 63(3):523-544.

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

Dekker, T., Koster, P., and Brouwer, R., 2014. Changing with the Tide: Semiparametric Estimation of Preference Dynamics. *Land Economics*, 90(4):717-745.

Directorate General of the State Forests, 2011. Bialowieza primeval forest - restoration. Warsaw.

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.

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.
McFadden, D., 2001. Economic Choices. *The American Economic Review*, 91(3):351-378.
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.
Navarro, L. M., and Pereira, H. M., 2012. Rewilding Abandoned Landscapes in Europe. *Ecosystems*, 15(6):900-912.
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*.
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.
Swait, J., and Louviere, J., 1993. The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models. *Journal of Marketing Research*, 30(3):305-314.
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.
Wesołowski, T., 2005. Virtual Conservation: How the European Union is Turning a Blind Eye to Its Vanishing Primeval Forests. *Conservation Biology*, 19(5):1349-1358.
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

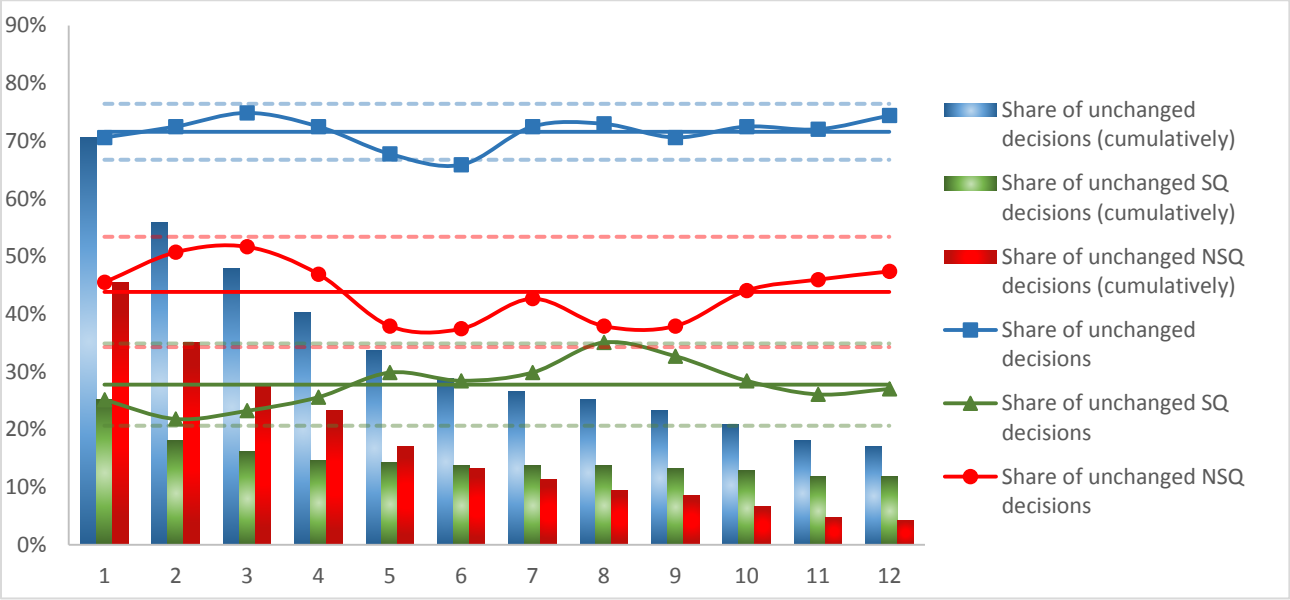


Figure A1. Dynamics of the number of decision changes in different choice tasks for respondents who did not participate in the second survey (set A vs. B; n = 211)

Table A1. Kruskal-Wallis test statistics.

	Share of unchanged decisions	Share of unchanged SQ decisions	Share of unchanged NSQ decisions
1	0.2924	1.8592	0.5748
2	0.1092	2.4483	2.8132*
3	0.0145	2.3435	1.5943
4	0.8821	3.5247*	0.8615
5	3.4101*	3.8423**	0.0584
6	0.4526	1.0896	0.1250
7	0.0848	2.5477	1.6712
8	0.1855	2.4935	4.2951**
9	0.7054	2.3143	0.5708
10	0.1917	1.2374	2.1369
11	0.3808	1.8733	0.5118
12	0.2426	0.3290	0.0092