Environ Resource Econ
DOI 10.1007/s10640-016-0045-z

# Choice Consistency and Preference Stability in Test-Retests of Discrete Choice Experiment and Open-Ended Willingness to Pay Elicitation Formats

**Roy Brouwer**[1,2] · **Ivana Logar**[2] · **Oleg Sheremet**[3]

Accepted: 16 June 2016


**Abstract** This study tests the temporal stability of preferences, choices and willingness to pay (WTP) values using both discrete choice experiment (DCE) and open-ended (OE) WTP elicitation formats. The same sample is surveyed three times over the course of two years using each time the same choice sets. Choice consistency is positively correlated with choice certainty and choice complexity. The impact of choice complexity fades away in time, most likely as a result of learning and preference refinement. Although the OE WTP values remain stable over a time period of 2 years as in previous stated preference studies, DCE based WTP measures differ significantly, suggesting their use in benefits transfer may be limited.

**Keywords** Preference stability · Choice consistency · Discrete choice experiment · Test-retest · Open-ended WTP

## 1 Introduction

The results of stated preference (SP) surveys are commonly used to estimate non-market values associated with proposed environmental changes at alternate study sites. As it is common for the source study and benefit transfer applications to take place several years apart, it is at least implicitly assumed that underlying preferences are temporally stable (Brouwer 2006) and choices are consistent (Schaafsma et al. 2014). The validity and reliability of these assumptions have been tested in different ways and over different time periods in SP studies, varying from just a few weeks or months to several years, using either the same

✉ Roy Brouwer
rbrouwer@uwaterloo.ca

1 The Water Institute, Department of Economics, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada

2 Swiss Federal Institute of Aquatic Science and Technology (Eawag), Dübendorf, Switzerland

3 Department of Geography and Sustainable Development, University of St Andrews, St Andrews, United Kingdom

Published online: 08 July 2016

or different samples (e.g., Whitehead and Hoban 1999; Berrens et al. 2000; Brouwer and Bateman 2005; Brouwer 2012; Fetene et al. 2014). Based on an overview of the existing test-retest contingent valuation (CV) literature, McConnell et al. (1998) concluded that SP and corresponding willingness to pay (WTP) values are stable over a time period of two years.

Discrete choice experiments (DCEs) have the advantage that through repetition respondents are expected to be capable of making more precise and consistent decisions, because they learn about the survey format, the associated hypothetical market environment and their own preferences (List 2003). Hoeffler and Ariely (1999) show that preference stability is positively correlated with choice experience (single versus repeated choice) and choice effort (easy versus difficult choice). This is fundamentally different from the preference elicitation formats employed in the early SP literature, where respondents are directly asked for their WTP, either through an open-ended (OE) or dichotomous choice (DC) WTP question. The OE question has been shown to produce more conservative WTP estimates than DC formats, but respondents have been reported to experience more uncertainty when answering OE than DC WTP questions (Bateman et al. 1995).

The empirical evidence for DCEs is much more limited than for other SP elicitation formats and mainly related to health care (for an overview see, for example, Mørbak and Olsen 2014). The results for DCEs confirm that WTP values are stable over short time periods (2 weeks to a few months). In the environmental valuation domain, only two published studies administered the same choice set sequence again after approximately one year to the same sample of respondents. Liebe et al. (2012) and Schaafsma et al. (2014) find similar results in terms of choice consistency and WTP estimates. However, Schaafsma et al. (2014) show that although WTP values remain stable, underlying preferences in the estimated choice models changed significantly over the one-year time period. Bliem et al. (2012) also tested the same choice set sequence over a one-year time period, but used different samples. They too found no significant differences in WTP.

This paper differs from the existing literature in that it tests and compares the temporal stability of SP and WTP values applying both a DCE and OE WTP elicitation format for an identical environmental change at three instead of two points in time. The same sample of respondents is offered the same sequence of choice sets and OE WTP questions exactly 6 and 24 months after they participated in the original survey. This provides more insight in preference dynamics than a test-retest at two points in time. The main objective of the study is to determine to what extent (1) choices in the same DCE are consistent, (2) the underlying indirect utility functions and WTP estimates derived from the DCE are stable, and (3) WTP estimates obtained from the OE WTP elicitation format are stable across the three surveys. Choice consistency is defined in this study in the strictest sense possible, i.e. as a respondent's choice for the same alternative in the identical choice set, either within (e.g. Brouwer et al. 2010) or between surveys (e.g. Schaafsma et al. 2014). Based on theory and the available empirical evidence, our a priori expectation is that, unless major changes occurred in the socio-economic situation of respondents (for which we control), preferences are stable, choices consistent and hence both the DCE and OE WTP estimates remain the same over a period of 2 years. Note that this does not necessarily mean that the OE and DCE based WTP values have to be the same for the same environmental change. The empirical evidence regarding the equality of welfare estimation derived from these two elicitation formats is mixed (e.g. Foster and Mourato 2003; Mogas et al. 2006). Our main interest here is to test the temporal stability of preferences and WTP values for each of the two elicitation formats.

## 2 Econometric Modeling Framework and Hypothesis Testing

### 2.1 Econometric Modeling Framework

Preferences for the policy alternatives chosen in the DCE are modeled in terms of McFadden's (1974) Random Utility Model (RUM), allowing for a separation of utility ($U_{ij}$) into a deterministic ($V_{ij}$) and stochastic part ($\varepsilon_{ij}$) [Eq. (1)]. The deterministic component of alternative $j$ can be specified as a linear function of its attributes ($\boldsymbol{X}$) and, possibly, other explanatory variables ($\boldsymbol{Z}$), where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the vectors of parameters associated with $\boldsymbol{X}$ and $\boldsymbol{Z}$ respectively:

$$U_{ij} = V_{ij} + \varepsilon_{ij} = \boldsymbol{\beta} \boldsymbol{X}_{ij} + \boldsymbol{\gamma} \boldsymbol{Z}_{ij} + \varepsilon_{ij} \tag{1}$$

The probability that alternative $j$ is chosen by an individual $i$ from a choice set sequence $\boldsymbol{C}$ is:

$$P_i\,(j|\,\boldsymbol{C}) = P\left[(V_{ij} + \varepsilon_{ij}) \geq (V_{ik} + \varepsilon_{ik})\ \forall k \in \boldsymbol{C},\ j \neq k\right] \tag{2}$$

In order to estimate Eq. (2), assumptions have to be made about the distribution of the error terms. Mixed distributions are most common nowadays in the DCE literature. Mixed logit models are flexible as they allow the model coefficients to vary across individuals, account for correlation across alternatives and generate unrestricted substitution patterns (McFadden and Train 2000). By enabling coefficients to vary randomly, the model is able to account for unobserved preference heterogeneity between individuals.

In the mixed logit model, the probability of observing a sequence of choices $y_i$ for an individual $i$ is conditional on the parameter vector $\boldsymbol{\beta}_i$. This vector of parameters is estimated as the product of conditional probabilities of all choice sets $t$ = (1,…,T) presented to the respondent:

$$P\left(y_i|\boldsymbol{\beta}_i\right) = \prod_{t=1}^{T} \frac{e^{(\boldsymbol{\beta}_i \mathbf{X}_{ij} + \boldsymbol{\gamma} \mathbf{Z}_{ij})}}{\sum_{k \in \boldsymbol{C}} e^{(\boldsymbol{\beta}_i \mathbf{X}_{ik} + \boldsymbol{\gamma} \mathbf{Z}_{ik})}} \tag{3}$$

Since the researcher does not observe $\boldsymbol{\beta}_i$, Eq. (3) is integrated over all possible values of $\boldsymbol{\beta}_i$ using their density function $f\,(\boldsymbol{\beta}|\boldsymbol{\theta})$:

$$P\,(y_i|\boldsymbol{\theta}) = \int_{-\infty}^{+\infty} P\left(y_i|\boldsymbol{\beta}_i\right) f\,(\boldsymbol{\beta}|\boldsymbol{\theta})\,d\beta \tag{4}$$

The probability of a sequence of choices for any individual is therefore conditional on the parameters of the density distribution, which is typically assumed to be normal: $f(0,\ \sigma^2)$. The integral in Eq. (4) cannot be solved analytically and requires simulation. The probability is approximated through a simulated maximum likelihood, which generates draws from distributions with given means and standard deviations. Here we will use 1000 Halton draws to ensure the efficiency of the maximum likelihood simulation procedure and model stability (Bhat 2001).

### 2.2 Hypothesis Testing Procedure

Following the main objectives, we test four hypotheses in this study. The first hypothesis relates to choice consistency in the DCE, which we define here as the individual respondent's choice $y_i$ of the same alternative $j$ in choice set $t$ across surveys:

$$H_0^1 : y_{ijt,test} = y_{ijt,retest} \tag{5}$$

To test this first hypothesis, we compare choices for each respondent across choice sets at three points in time, which gives us the proportion of consistent choices. The non-parametric Sign test is applied to test the equality of choices.

The second hypothesis relates to the stability of the underlying indirect utility function presented in Eq. (1). The estimated preference parameters $\boldsymbol{\beta}$ and scale parameters $\lambda$ in the utility function are expected to be the same across the surveys:

$$H_0^{2a} : \boldsymbol{\beta}_{i,test} = \boldsymbol{\beta}_{i,retest} \tag{6}$$

$$H_0^{2b} : \lambda_{test} = \lambda_{retest} \tag{7}$$

The scale parameter $\lambda$ is inversely related to the variance of the error term $\varepsilon$ in Eq. (1) (Louviere et al. 2000). Scale increases (variance decreases) are typically associated with refined and more accurate choices between alternatives (Holmes and Boyle 2005). In order to test equality of preference and scale parameters in the test and retests, we follow the Swait and Louviere (1993) two-step test procedure. First, two separate models are estimated independently based on the test and retest choice data and compared to a pooled model for the two samples together. For this latter pooled model, a grid search for the scale parameter is performed to optimize the log-likelihood function under different relative scale adjustments for one of the two datasets to keep the scale parameter constant between datasets. A Likelihood Ratio (LR) test is subsequently performed to see if restricting the preference parameters to be equal for the two datasets results in a significantly different model fit: $\widehat{\boldsymbol{\beta}}_{\text{test}} = \widehat{\boldsymbol{\beta}}_{\text{retest}}$. If the null-hypothesis of equal preference parameters is rejected, there may be differences between the two datasets. It is however not possible to attribute these to differences in either the preference or scale parameters or both. If the null hypothesis cannot be rejected, then the LR test is applied again to test the hypothesis that there is no significant difference in the scale parameters between the two datasets: $\hat{\lambda}_{\text{test}} = \hat{\lambda}_{\text{retest}}$.

The third hypothesis tests the equality of the estimated mean WTP values across the three surveys for each elicitation format:

$$H_0^{3a} : WTP_{test}^{\text{DCE}} = WTP_{retest}^{DCE} \tag{8}$$

$$H_0^{3b} : WTP_{test}^{OE} = WTP_{retest}^{OE} \tag{9}$$

The DCE based WTP values are derived from the same choice models applied to test the second hypothesis, but estimated in WTP space (e.g. Daly et al. 2012). The third hypothesis is tested using the Wald test.

Finally, in view of the fact that the OE WTP always follows the DCE, we test for possible ordering effects by using split samples, one receiving the OE WTP question after and one before the DCE. This gives us our fourth and final hypothesis:

$$H_0^4 : WTP^{OE\ before\ DCE} = WTP^{OE\ after\ DCE} \tag{10}$$

In this case, equality of mean WTP will be tested using the non-parametric Mann-Whitney test given that the WTP values are elicited independently from each other in split samples.

## 3 Case Study

The increasing contamination of freshwater systems with thousands of chemical compounds from pharmaceuticals, personal care products, pesticides, and chemicals used in industry is a growing environmental concern worldwide. A common name for these chemicals in

water bodies at low concentrations is micropollutants (MPs). Despite their low concentrations, there are indications that MPs have potentially adverse impacts on aquatic ecosystems (Kidd et al. 2007). However, little is known about their implications for the environment and human health. New wastewater treatment technologies have been developed, which are able to remove up to 80 percent of MPs from wastewater (Hollender et al. 2009). This new technology is very costly. This study explores public preferences and WTP for upgrading wastewater treatment plants with the aim of removing MPs from water bodies and hence reduce their possible environmental and health risks.

The environmental risk attribute in the DCE in this study reflects the impact MPs are expected to have on aquatic flora and fauna. Levels of the potential environmental risk attribute were defined based on the number of MPs that exceed their environmental quality standards downstream from wastewater treatment plants in Switzerland. A detailed description is provided in Logar et al. (2014). Current environmental risk levels across the whole of Switzerland vary from low to high and were visualized in a map that was shown to respondents in the survey.

Given that Switzerland is a federation where distribution of drinking water and sanitation of wastewater fall within the competence of the cantons, people were offered the possibility to choose between removal of MPs in wastewater treatment plants at national or regional (canton) level. Hence, the second attribute captures respondents' preferences for the spatial scale of a reduction in the potential environmental risk.

The third attribute in our DCE measures the expected number of years necessary before scientific knowledge about the impacts of MPs on human health would become available by extra investments in relevant scientific research. Such knowledge would support policy and decision-making to upgrade wastewater treatment plants for the sake of safeguarding human health. Based on expert judgment, it was expected to take at least another 20 years in the current situation before a well-informed decision can be made about the health risks involved. Respondents were explained that investing more in scientific research could reduce this to 15, 10 or 5 years.

The possible environmental and health benefits had to be traded off against an increase in the household's annual water bill, varying between 10 and 150 Swiss Francs (CHF). A main effects fractional factorial design was used. Out of several design variants generated by the software Sawtooth CBC we selected the one with the lowest reported D-error and checked for the presence of strictly dominant alternatives. This resulted in 38 different versions, each consisting of six choice sets. A version was randomly assigned to respondents. The six choice sets included two hypothetical policy scenarios and the option to choose none of the two and stay with the status quo. Respondents were informed that this third option meant that their water bill would not increase, current potential environmental risk levels stay the same and would not be reduced to a low potential risk level, and that it would take at least another 20 years before more information would be available about the human health risks.

The design was pre-tested in three rounds in April and May 2012 with the help of a hired marketing company specialized in public surveys and communication. The first two pre-tests consisted of 76 face-to-face interviews and led to several changes in the formulation of questions and explanation related to the DCE. The third and final pre-test was conducted online among a sample of 122 respondents.

The attributes in the DCE were presented to respondents both as pictograms and as text. An example of a choice set is presented in Fig. 1. After each choice set, respondents were asked to rate their level of certainty regarding the choice they made on a scale from 0 to 10, where 0 indicates not certain at all and 10 completely certain.

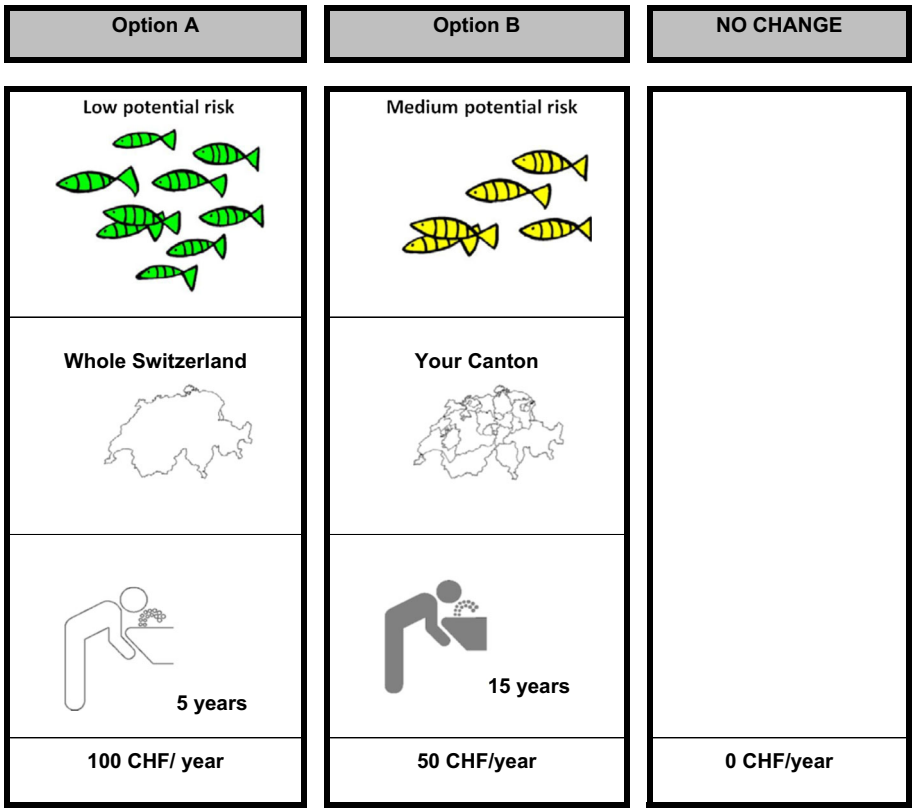


**Fig. 1** Example of a choice card

Following the DCE, all respondents were asked in an OE WTP question what they would be willing to pay maximum for the policy scenario where the potential environmental risk is reduced to the lowest level for the whole country. This is a commonly applied procedure to allow for a direct comparison between the WTP values derived from the DCE and OE WTP elicitation formats for a particular policy scenario (e.g. Hynes et al. 2011). Typically, the OE WTP question generates lower values than discrete choice questions in the early SP literature (e.g. Bateman et al. 1995). An OE WTP question was used instead of a dichotomous choice WTP question because this came out as the preferred elicitation format based on pretesting. Respondents reported to feel fatigue when asked to go through another series of discrete choices related to their WTP after the DCE.

The web-based survey was programmed in the software Sawtooth CBC. The first survey was administrated over the internet in May 2012. It was conducted in the 20 German-speaking cantons of Switzerland, which form the majority of the country's 26 cantons. The survey sample was set up to be representative for this part of Switzerland. Respondents above the age of 18 were randomly drawn from a consumer panel provided by the marketing agency ISO-PUBLIC, who has access to 35 thousand Swiss households. The panel is compiled through face-to-face and computer assisted telephone interviewing to ensure precise and representative sampling based on household profiles related to socio-demographic characteristics, profession, personal interests, internet knowledge, finance, health status, travel and vacation behavior.

A total of 1000 respondents completed the online questionnaire (response rate of 25 %). The same sample of 1000 respondents was contacted again exactly 6 and 24 months after they participated in the first survey. The respondents in the two retests were asked a few additional questions to see if anything had changed in their opinion or their personal circumstances that could have influenced their choices. They received exactly the same choice sets in the same order as they had in the original survey.

## 4 Results

### 4.1 Sample Characteristics and Self-Selection Bias

Out of the original 1000 respondents, 304 completed the online survey in the first retest and another 245 respondents in the second retest. These follow-up response rates (30 and 25 %) are lower than the follow-up response rate of 48 % found in Liebe et al. (2012), but comparable to the 28 % in Schaafsma et al. (2014). A possible reason for this is that Liebe et al. (2012) only approached those respondents who replied positively to the question whether they agree to be interviewed again after the original survey, which may have induced self-selection bias. In our study, the characteristics of the respondents who participated in the two retests are similar to those who only participated in the first survey (i.e. non-repeaters). This is shown in Table 1. A slightly higher share of male and retired respondents who are more often a member of an environmental protection organization completed the survey the second and third time. Slightly less non-repeaters have a university degree, while slightly more are self-employed or full-employed. No significant differences can be detected between the repeaters and non-repeaters in terms of education level and average household income. Less than 10 % of those who participated in the first retest (n=26) reported a change in their personal circumstances over the six-month time period compared to when they participated in the original survey (e.g. household size, employment status, income), while this share is less than 25 % ($n = 56$) in the second retest 2 years later.

Possible self-selection bias was tested using a Heckman Full Information Maximum Likelihood model in Stata version 13 (see Table 2). Two functions are estimated simultaneously, one specifying the probability that a respondent agrees to participate in the retest and one specifying the OE WTP bid function. The Heckman procedure tests whether participation in the retest is random between the sample of repeaters and non-repeaters. If non-random, the factors that determine participation in the retest may be correlated to the factors that determine the WTP values, resulting in a violation of the assumption of independent and identically distributed errors (e.g. Messonnier et al. 2000).

As shown in Table 2, we find a significant negative correlation ($\rho$) between the decision to participate in the retest and OE WTP, indicating that the estimated WTP function suffers from selection bias. The negative coefficient for both $\rho$ and the selection coefficient $\lambda$ suggests that unobserved factors that make participation more likely tend to be associated with lower WTP. In the selection models, male respondents are more likely to participate in the two retests. Respondent age and the importance respondents attach to water quality are also significant determinants of the likelihood of participating in the first retest, while membership of an environmental organization has a significant positive effect on the decision to participate in the second retest. In the OE WTP models, gender and income only play a significant role in the second retest. We also tested the possible impact of a number of survey characteristics, such as the recorded amount of time it took respondents to complete the surveys, the self-

**Table 1** Socio-demographic characteristics of the test and retest samples and the sample of non-repeaters

| Characteristic | Original sample (May 2012) | Repeaters | | Non-repeaters* |
|---|---|---|---|---|
| | | First retest (November 2012) | Second retest (May 2014) | |
| Share male (%) | 56.9 | 63.2 | 63.7 | 53.8 |
| Average age[a] | 52.4 (25–75) | 54.2 (25–74) | 53.6 (25–75) | 51.5 (26–75) |
| Average household size[a] | 2.7 (1–9) | 2.5 (1–7) | 2.6 (1–7) | 2.7 (1–9) |
| Average number of children[a] | 0.7 (0–7) | 0.6 (0–6) | 0.6 (0–6) | 0.7 (0–7) |
| Share primary school degree only (%) | 1.3 | 0.7 | 0.8 | 1.7 |
| Share secondary school degree (%) | 63.8 | 62.8 | 63.3 | 64.1 |
| Share university degree (%) | 12.1 | 14.5 | 15.1 | 10.3 |
| Share unemployed (%) | 1.3 | 1.6 | 1.2 | 1.2 |
| Share self-employed (%) | 8.9 | 7.6 | 8.2 | 9.6 |
| Share full-employed (%) | 37.7 | 36.5 | 34.3 | 39.0 |
| Share housewife/man (%) | 5.4 | 4.9 | 5.7 | 5.0 |
| Share retired (%) | 19.9 | 25.0 | 24.5 | 17.2 |
| Average monthly income (CHF)[b] | 8641 (8250) | 8797 (8250) | 8519 (8250) | 8597 (8250) |
| Member of an environmental organization (%) | 45.1 | 50.3 | 52.2 | 40.5 |
| Number of respondents | 1000 | 304 | 245 | 582 |

[a] Min-max range between brackets
[b] Median between brackets
* Respondents who did not participate in the first and second retest

reported importance they attach to each of the choice attributes, respondent authorization to add their stated WTP to their next water bill as an indicator of hypothetical bias or the perceived credibility of the policy scenarios. None of these characteristics are statistically significant and are therefore not presented here.

Possible selection bias was also examined by comparing the estimated choice models between the group of respondents who participated in the test and retests. Applying the Swait and Louviere (1993) test procedure, we are unable to reject the null hypothesis of equal preference parameters between those respondents who participated in the original survey and the first retest and those respondents who participated in the original survey and the second retest. Only a significant difference in scale parameter can be detected between the original sample and second retest. Hence, the retests do not significantly affect the preference

**Table 2** Estimated sample selection regression models for the first and second retest

| Variable | Variable definition | First retest | | Second retest | |
|---|---|---|---|---|---|
| | | Coefficient estimate | Std. error | Coefficient estimate | Std. error |
| *Selection model* | | *Participation in first retest = 1* | | *Participation in second retest = 1* | |
| Constant | | −1.150 | 0.720 | −0.347 | 0.749 |
| Gender | Dummy 1=male respondent | 0.194** | 0.086 | 0.229*** | 0.091 |
| Age | Years | 0.008** | 0.004 | 0.002 | 0.004 |
| Household income | Natural log 1000 CHF/month | −0.008 | 0.077 | −0.081 | 0.081 |
| Member environmental organization | Dummy 1 = yes | 0.036 | 0.084 | 0.200** | 0.088 |
| Importance attached to water quality | Dummy 1 = very important | 0.176*** | 0.055 | 0.054 | 0.069 |
| Concern about MPs in water | Dummy 1 = very concerned | 0.002 | 0.070 | 0.100 | 0.091 |
| *OE WTP model* | | *OE WTP in first retest* | | *OE WTP in second retest* | |
| Constant | | 6.320*** | 1.525 | 3.622** | 1.577 |
| Gender | Dummy 1 = male respondent | −0.282 | 0.182 | −0.477*** | 0.193 |
| Age | Years | −0.010 | 0.008 | −0.005 | 0.008 |
| Household income | Natural log 1000 CHF/month | 0.088 | 0.161 | 0.386** | 0.169 |
| Member environmental organization | Dummy 1 = yes | −0.067 | 0.176 | −0.035 | 0.186 |
| *Model summary statistics* | | | | | |
| $\sigma$ | | 2.024*** | 0.111 | 1.908*** | 0.140 |
| $\lambda$ | | −1.970*** | 0.116 | −1.785*** | 0.159 |
| $\rho$ | | −0.973*** | 0.006 | −0.936*** | 0.018 |
| $LR\ \chi^2\ test\ (\rho=0)$ | | 103.93 | $p$ <0.001 | 34.01 | $p$ <0.001 |
| *LL* | | −1041.878 | | −913.256 | |
| *Number of observations (number of non-repeaters)* | | 1000 (696) | | 1000 (755) | |

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$

coefficients in the estimated choice models. As a result, no significant difference can be found between the estimated mean WTP welfare measures for respondents who participated in the test only and those who participated in the retests.[1]

### 4.2 Choice Consistency

Choice consistency was measured by comparing respondent choices across each choice set in the test and two retests and counting the number of times the same alternative was chosen in each choice occasion. Choice shares for the three alternatives across choice sets and surveys are shown in Fig. 2 for the original sample and the samples of repeaters. The choice shares do not differ much between the test and two retests and follow the same pattern. The first alternative is chosen slightly more often than the second alternative in the second choice set in the three surveys. The opt-out share is slightly higher in the first retest (13.9 %) than in the original survey (10.8 %) and second retest (9.7 %). The share of respondents who consistently chose the opt-out (i.e. six times) is also highest in the first retest (7.6 %) compared to the original survey (3.9 %) and the second retest (2.4 %), but is overall relatively low across all three surveys. This is, among others, due to extensive pre-testing of the DCE.

In order to test the first hypothesis, the share of consistent choices among repeaters across the six choice sets is compared in Table 3 for the three surveys. The choice consistency rate at choice set level varies between 56 and 66 % across the 3 surveys. Sixty-three percent of the choices are consistent when comparing choice behavior between the original survey and the first retest. This is slightly higher than the 59 % found in Liebe et al. (2012) and the 57 % found in Schaafsma et al. (2014). Choice consistency is, as expected, lowest after 24 months when comparing choices in the original survey and the second retest (59 %), but still directly comparable to the findings reported in the existing literature over a time period of 12 months. The first and second retest are 18 months apart from each other and produce a slightly higher choice consistency of 61 %.

One in five respondents consistently chose exactly the same six alternatives, i.e. made identical choices across all six choice sets, in the first retest and the original survey (Table 4). This is halved when comparing the original survey and the second retest and the first and second retest. Between three and six percent was completely inconsistent, choosing a different alternative in each choice set in the retests compared to the original survey. Although the share of inconsistent choices is considerable in the two follow-up surveys (37–41 %), the non-parametric Sign test is unable to reject the null hypothesis of equal choices at choice set level between test and retests.

Choice consistency was also measured within the same survey in the original survey and the second retest. A sub-sample of 250 respondents in the original survey was shown 7 choice sets where the first set was repeated without informing respondents either as the fourth, fifth, sixth or seventh choice set. The position in the choice set sequence where the choice set would be repeated was randomized across respondents. In the original survey, 79.6 % of these 250 respondents chose the same alternative again at a later point in the choice sequence. Some variation was found in the consistency rate depending on the location in the choice sequence where the first choice set was repeated. When shown again as the fourth choice set, 82 % chose the same alternative again, while if shown again as the seventh choice set, 75 % chose the same alternative again. In the second retest the same consistency test was carried out once again with the same respondents, but this time with a much smaller sub-sample size of only 28 respondents. The consistency rate based on this small sample was higher, namely 90.6 %

[1] The test results are available from the authors.

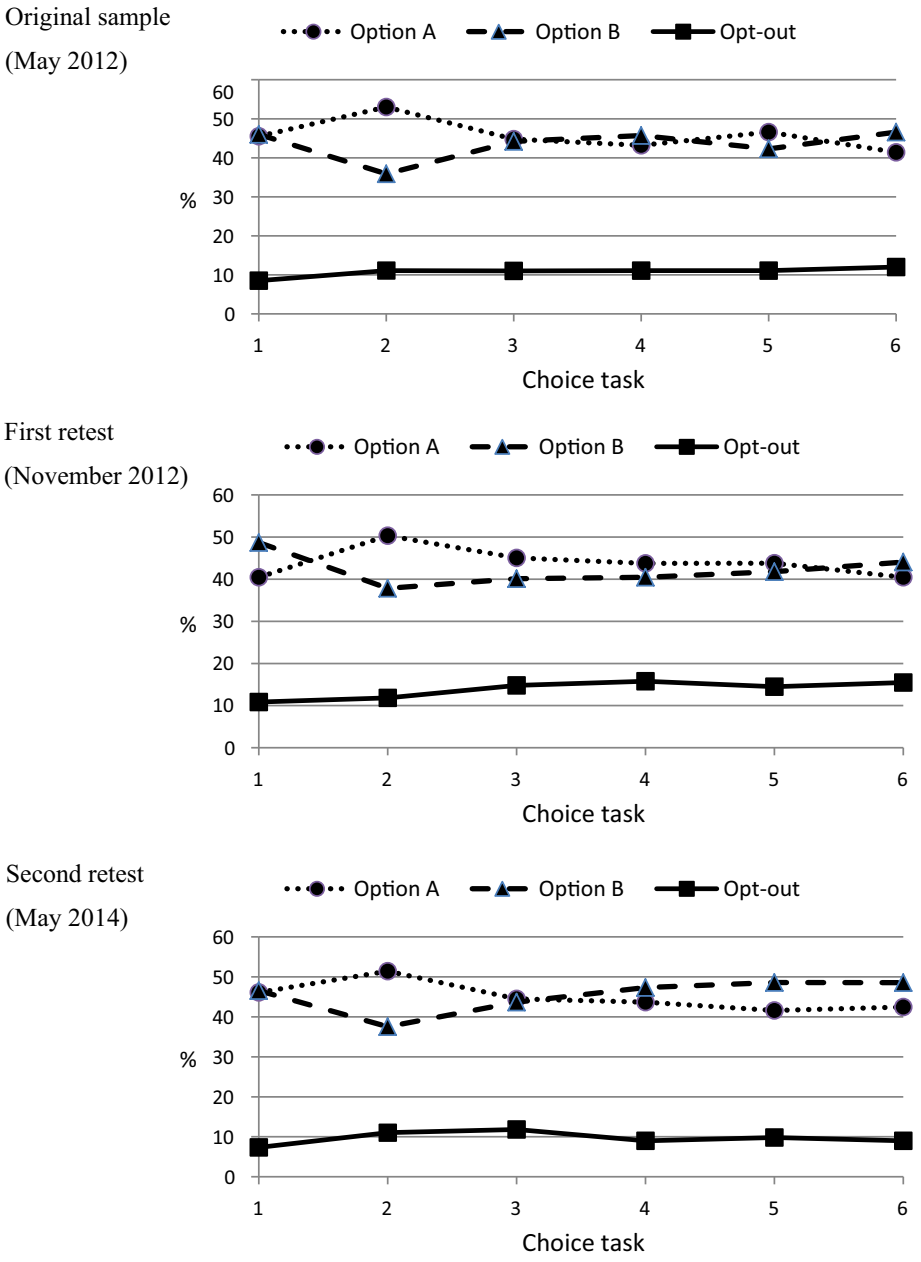


**Fig. 2** Choice shares per choice task in the test (*top*) and two retests (*middle* and *bottom*)

(varying between 71 and 100 % depending on where in the choice sequence the first choice set was repeated).

Finally, respondent choice consistency for the samples of repeaters was regressed on possible explanatory factors in a random effects probit model, accounting for the panel data

**Table 3** Repeater sample shares (%) consistently choosing the same policy alternatives in the same discrete choice experiment at different points in time

| Choice set | Original survey and first retest ($n = 131$) | | Original survey and second retest ($n = 131$) | | First and second retest ($n = 131$) | |
|---|---|---|---|---|---|---|
| | Consistent choices (%) | Sign test statistic | Consistent choices (%) | Sign test statistic | Consistent choices (%) | Sign test statistic |
| 1 | 66.4 | 0.754 ($p = 0.451$) | 64.1 | 0.001 ($p = 1.000$) | 61.1 | −0.840 ($p = 0.401$) |
| 2 | 61.8 | 0.707 ($p = 0.480$) | 55.7 | 1.707 ($p = 0.088$) | 61.1 | 0.560 ($p = 0.575$) |
| 3 | 64.1 | 0.001 ($p = 1.000$) | 60.3 | −0.416 ($p = 0.677$) | 58.8 | −0.408 ($p = 0.683$) |
| 4 | 61.8 | 0.424 ($p = 0.671$) | 58.8 | 0.680 ($p = 0.496$) | 61.8 | 0.141 ($p = 0.888$) |
| 5 | 60.3 | 1.248 ($p = 0.212$) | 58.5 | 0.001 ($p = 1.000$) | 64.1 | −1.167 ($p = 0.243$) |
| 6 | 60.3 | 0.139 ($p = 0.890$) | 58.0 | −0.539 ($p = 0.590$) | 58.8 | −0.408 ($p = 0.683$) |
| Total | 62.5 | | 58.9 | | 60.9 | |

Sign test is the non-parametric related samples Sign test

**Table 4** Number of times repeaters consistently choose the same policy alternatives in the same discrete choice experiment at different points in time (%)

| Number of identical choices | Original survey and first retest ($n = 131$) | Original survey and second retest ($n = 131$) | First and second retest ($n = 131$) |
|---|---|---|---|
| 0 | 6.1 | 3.1 | 5.3 |
| 1 | 6.1 | 8.4 | 7.6 |
| 2 | 10.7 | 17.6 | 12.2 |
| 3 | 19.1 | 19.8 | 10.7 |
| 4 | 19.8 | 16.8 | 28.2 |
| 5 | 18.3 | 22.9 | 26.7 |
| 6 | 19.8 | 11.5 | 9.2 |
| Total | 100.0 | 100.0 | 100.0 |

structure (6 choices per individual in each survey). These explanatory factors are related to respondents' socio-demographic characteristics (e.g. sex, age, income, education level), their attitudes (e.g. membership of an environmental organization, concern about MPs), their perception of the survey instrument and the choices they were asked to make (e.g. credibility, choice certainty and choice realism, which is based on authorizing the marketing company to add the stated bid amount to their next water bill), their choice behavior (e.g. the time needed to complete the survey, whether respondents consistently chose the opt-out), and the characteristics of the choice experiment (the standardized difference between the levels of the alternatives in the DCE indicating design complexity).

The results in Table 5 show that the probability of repeaters exhibiting consistent choice behavior depends significantly on the self-reported choice certainty. The more certain a respondent is about a choice, the higher the probability of choosing the same alternative again in the next survey. The inclusion of self-reported choice certainty in choice models could arguably result in potential endogeneity bias when modelling choice behavior (e.g. Dekker et al. 2016). Less certain respondents are expected to make more random choices and choose more frequently the opt-out alternative. Such a relationship between opt-out choices and choice certainty was however not found in this study. Another remarkable finding is that, as expected, the bigger the difference between the attribute levels of two alternatives, the higher the likelihood of choosing the same alternative again in the next survey.[2] However, this effect fades away in time: the difference is highly significant when comparing choices over the shortest time period of 6 months (original survey and first retest), less significant when comparing choices over 18 months (first and second retest), and not significant when comparing the choices over a time period of 24 months (original survey and second retest). Moreover, respondents who clearly understood what they were being asked to pay for, a debriefing question at the end of the survey, exhibit significantly higher choice consistency. If the information about the environmental good for which respondents were asked to pay was considered clear, this also resulted in more consistent choices.

The influence of other explanatory variables is very limited. None of the socio-demographic or attitudinal characteristics are statistically significant, except membership

[2] Differences were calculated by subtracting the levels of the choice attributes in the non-chosen alternative from the levels of the chosen alternative, standardizing these differences per attribute and aggregating them linearly.

**Table 5** Random effects probit regression model explaining consistent choice behavior in the same discrete choice experiment at different points in time

| Explanatory factor | Original survey and first retest ($n = 304$) | | Original survey and second retest ($n = 245$) | | First and second retest ($n = 131$) | |
|---|---|---|---|---|---|---|
| | Coefficient estimate | Standard error | Coefficient estimate | Standard error | Coefficient estimate | Standard error |
| Constant | −0.807 | 0.432 | −0.249 | 0.383 | −0.395 | 0.590 |
| Respondent gender (1= male) | −0.195 | 0.123 | −0.071 | 0.113 | −0.133 | 0.178 |
| Respondent age (years) | 0.0004 | 0.005 | −0.007 | 0.005 | −0.004 | 0.008 |
| Respondent household size (persons) | −0.019 | 0.051 | −0.007 | 0.047 | −0.080 | 0.070 |
| University degree (1= yes) | 0.224 | 0.168 | 0.190 | 0.155 | −0.067 | 0.224 |
| Household income (1000 CHF/month) | 0.009 | 0.014 | −0.004 | 0.013 | 0.019 | 0.018 |
| Member environmental organization (1= yes) | 0.364*** | 0.116 | 0.103 | 0.106 | 0.400*** | 0.159 |
| Living distance from water (1 ≤ 1 km) | 0.517 | 0.112 | −0.071 | 0.104 | 0.137 | 0.160 |
| Clear what being asked to pay for (1= yes) | 0.107 | 0.154 | 0.333** | 0.141 | −0.081 | 0.216 |
| Authorize to add amount to water bill (1= yes) | 0.061 | 0.154 | 0.037 | 0.125 | −0.035 | 0.200 |
| Choice certainty (0–10) | 0.093*** | 0.024 | 0.074*** | 0.022 | 0.092*** | 0.031 |
| Credibility of the presented information (1= very credible) | 0.192 | 0.176 | 0.296** | 0.153 | 0.443** | 0.225 |
| Concern about MPs in water (1= very concerned) | 0.167 | 0.168 | −0.155 | 0.141 | 0.017 | 0.221 |
| Standardized difference between alternatives (0–1) | 0.693*** | 0.150 | −0.019 | 0.090 | 0.232* | 0.126 |
| Consistent opt-out voter (1= yes) | 6.897 | 137.076 | 6.432 | 199.450 | 6.515 | 181.876 |
| Average time to complete survey (minutes) | 0.003 | 0.004 | −0.001 | 0.005 | 0.013 | 0.019 |
| $\sigma$ | 0.734*** | 0.062 | 0.563*** | 0.062 | 0.614*** | 0.088 |
| $\rho$ | 0.350*** | 0.039 | 0.241*** | 0.040 | 0.274*** | 0.057 |
| Log-likelihood | −1050.018 | | −928.329 | | −472.049 | |
| Wald chi-square (15 d.o.f.) | 60.65*** | | 33.75*** | | 27.33*** | |
| Number of observations | 304 | | 245 | | 131 | |

Response variable: consistent choice = 1, otherwise = 0
*, ** and *** denote $p < 0.1$, $p < 0.05$ and $p < 0.01$

of an environmental organization and the credibility of the information presented in the survey about MPs. Both characteristics influence choice consistency in a positive way in two of the three models.

### 4.3 Equality of Preference and Scale Parameters in the Estimated Choice Models

In order to test the second hypothesis, random parameters logit (RPL) choice models are estimated in preference space in NLOGIT 5 for the full samples and the sample of respondents who participated in all three surveys ($n = 131$). Note that the latter is a sub-sample of 'pure' repeaters. As the results for the two samples are the same, we present and discuss here the models for the full samples in Table 6. The first model is based on the choices of 1000 respondents in the original survey, the second model on the choices of 304 respondents in the first retest and the third model on the choices of 245 respondents in the second retest.

Contrary to Liebe et al. (2012), we are unable to detect a significant error component in the estimated models in this study. Like Liebe et al. (2012), we present here the choice models including their choice attributes only. This is justified by the fact that we compare models across the same respondents, and we were unable to find significant differences between respondents who indicated that their socio-economic situation had changed and respondents who said it had not over the retest period. We are unable to replicate the results presented in Schaafsma et al. (2014) who found significant effects of respondent sex, household income and membership of an environmental organization on choice behavior. Only respondents' membership of an environmental organization had a significant positive effect on choice behavior in all the estimated choice models in this study.

All choice attributes in the models are highly significant at the one percent level, have the expected signs, and exhibit preference heterogeneity as can be seen from the highly significant standard deviation of the random coefficients. The positive sign for the ASC indicates that respondents prefer, ceteris paribus, the hypothetical alternatives over the current situation. Respondents have strong preferences for a national instead of cantonal policy to reduce the current medium and high potential risk levels to a low potential risk, and prefer the knowledge about the potential human health impacts of MPs to become available sooner rather than later (the negative sign indicates preferences for a shorter period of time). As expected, the price coefficient is highly significant and negative in the three models. The distribution of the random terms of the dummy attributes is uniform following recommendations by Hensher et al. (2005), while the best fit distribution for the availability of knowledge and the price attribute is normal. As can be seen from Table 6, the estimated coefficients and the standard deviations of the random parameters are similar across the three surveys. The marginal utility attached to a low environmental risk level slightly decreases from the original survey to the first retest and then declines further in the second retest. At the same time the marginal utility attached to the scale of policy implementation (whole country) slightly increases. Respondents' price sensitivity also increases across the three surveys.

We formally test the equality of the estimated choice models using the Swait and Louviere (1993) test procedure. The test outcomes show that the null hypothesis of equal preference parameters for respondents who participated in the two retests cannot be rejected at the 10 % significance level (first two rows in Table 7). The same results are found for respondents who participated in all three surveys ($n = 131$), except that the null hypothesis of equal preference parameters over a time period of 6 months is rejected at the 5 % level ($p = 0.026$). The estimated preference parameters are stable over a time period of 18 months (comparing the first and second retest) and 2 years (comparing the original survey with the second retest). When comparing the equality of scale parameters, we are unable to find a significant

**Table 6** Mixed logit discrete choice models estimated in preference space based on the same discrete choice experiment at different points in time

| | Original sample | | First retest | | Second retest | |
|---|---|---|---|---|---|---|
| | Coefficient estimate | Standard error | Coefficient estimate | Standard error | Coefficient estimate | Standard error |
| *Mean estimate random attribute parameters* | | | | | | |
| Alternative specific constant (ASC) | 5.806*** | 0.356 | 6.190*** | 0.718 | 5.884*** | 0.645 |
| Low environmental risk | 1.039*** | 0.070 | 1.028*** | 0.133 | 0.960*** | 0.150 |
| Whole country | 0.540*** | 0.058 | 0.609*** | 0.113 | 0.802*** | 0.138 |
| Availability of knowledge | −0.064*** | 0.006 | −0.072*** | 0.013 | −0.061*** | 0.013 |
| Price | −0.013*** | 0.001 | −0.015*** | 0.002 | −0.020*** | 0.003 |
| *Standard deviation random parameters* | | | | | | |
| Alternative specific constant (ASC) | 4.156*** | 0.333 | 5.207*** | 0.624 | 3.324*** | 0.569 |
| Low environmental risk | 2.100*** | 0.140 | 2.085*** | 0.263 | 2.395*** | 0.298 |
| Whole country | 1.606*** | 0.142 | 1.755*** | 0.254 | 2.054*** | 0.291 |
| Availability of knowledge | 0.084*** | 0.010 | 0.110*** | 0.018 | 0.084*** | 0.019 |
| Price | 0.028*** | 0.002 | 0.031*** | 0.003 | 0.034*** | 0.004 |
| *Model summary statistics* | | | | | | |
| Log-likelihood | −4280.920 | | −1303.524 | | −1030.179 | |
| LR chi-square (10 d.o.f.) | 4621.508*** | | 1400.691*** | | 1169.563*** | |
| McFadden $R^2$ | 0.351 | | 0.349 | | 0.362 | |
| Number of respondents | 1000 | | 304 | | 245 | |
| Number of observations | 6000 | | 1824 | | 1470 | |

*, ** and *** denote $p < 0.1$, $p < 0.05$ and $p < 0.01$

**Table 7** Results of the likelihood ratio tests of equal preference and scale parameters between the estimated discrete choice models at different points in time

| Sample $t$ | $N_t$ | Sample $r$ | $N_r$ | $LL_t$ | $LL_r$ | $LL_{pooled}$ ($\lambda_t \neq \lambda_r$) | LR-test (d.o.f. 11) | $H_0^a$: $\boldsymbol{\beta}_t = \boldsymbol{\beta}_r$ | $LL_{pooled}$ ($\lambda_t = \lambda_r$) | LR-test (d.o.f. 1) | $H_0^b$: ($\lambda_t = \lambda_r$) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Original survey | 304 | First retest | 304 | −1287.117 | −1305.354 | −2594.207 | 3.472 | $p = 0.983$ | −2594.370 | 0.327 | $p = 0.568$ |
| Original survey | 245 | Second retest | 245 | −1078.141 | −1032.311 | −2115.604 | 10.304 | $p = 0.503$ | −2114.194 | 2.820 | $p = 0.093$ |
| Original survey | 131 | First retest | 131 | −576.021 | −527.546 | −1115.279 | 21.772 | $p = 0.026$ | −1114.453 | 1.653 | $p = 0.198$ |
| Original survey | 131 | Second retest | 131 | −576.021 | −572.977 | −1156.903 | 15.273 | $p = 0.170$ | −1156.635 | 0.537 | $p = 0.464$ |
| First retest | 131 | Second retest | 131 | −527.546 | −572.977 | −1108.792 | 13.577 | $p = 0.257$ | −1107.312 | 2.961 | $p = 0.085$ |

*LL* log likelihood, *LR* likelihood ratio, *t* test, *r* retest, *d.o.f.* degrees of freedom

difference in the error variance between the test and retest models over a time period of 6, 18 and 24 months at the 5 % significance level.

### 4.4 Temporal Stability of the Estimated WTP Welfare Measures

In order to test the third hypothesis, the same test and retest choice models are estimated in WTP space in NLOGIT 5 to allow for direct comparison of welfare estimates across the three surveys. Table 8 shows the mean WTP estimates for the policy scenario of interest where the potential environmental risk is reduced to a low level in the whole country based on the test and retest models. The knowledge attribute is assumed to remain at its baseline level, implying that the knowledge about the impacts of MPs on human health will become available in 20 years from now. The welfare estimates for the retests are not adjusted for inflation given the fact that there was no inflation in Switzerland during the two-year period (according to Swiss Statistics, the inflation rates were −0.2 % in 2013 and 0.0 % in 2014).

Mean WTP derived from the DCE for the sample of respondents who participated in all three surveys ($n = 131$) initially drops significantly by 25 % over a time period of 6 months, but then slightly increases again by 10 % over the next 18 months. After two years, mean WTP is therefore 15 % lower than in the original survey. A similar result is found when comparing estimated mean WTP for respondents who participated in the original and the third survey ($n = 245$). Mean WTP is after 2 years the same for this group of respondents (CHF 86.7/ household/year) and for respondents who participated in all three surveys (CHF 86.8/ household/year). The latter estimate is less accurate due to the smaller sample size as can be seen from the wider 95 % confidence interval. The observed differences between the estimated mean WTP values in the test and retests for the sample who participated in all three surveys are statistically significant at the 5 % level. Hence, we reject the third hypothesis and conclude that the DCE based welfare measures are not stable over a time period of 6, 18 and 24 months. This result goes against findings in the existing literature that stated WTP in DCEs are stable over a time period of 12 months.

Following the DCE, respondents were asked for their maximum WTP for the same policy scenario in an OE question. In order to test for possible ordering effects (the fourth hypothesis in this study), we asked half of the respondents in the third survey the OE WTP question before the DCE. This generated a significantly higher mean WTP (CHF 157.3/household/year) at the 5 % level than when the same OE WTP question is asked after the DCE in the third survey (CHF 95.1/household/year).[3] We therefore reject the fourth hypothesis and account for these ordering effects in the third survey by only comparing mean WTP values which were elicited after the DCE ($n = 137$) with those from the first and the second survey. The results for the mean OE WTP values in the test and retests are also reported in Table 8.

The number of stated zero bids in the OE WTP question is very limited (between 2.5 and 5 % in the three surveys). Similar to the DCE results, we find no significant influence of any of the socio-economic respondent characteristics or the perception and attitudinal variables on the stated OE WTP values. Like Brouwer et al. (2015), we do find significant positive anchoring of the stated OE WTP values on the price level in the respondents' preferred choice alternative in the DCE's last choice set. This suggests that the two elicitation formats do not produce independent welfare measures. We use the mean predicted values to test the stability of OE WTP in time with the same Wald test in Stata.

Examining the mean OE WTP values in Table 8 across the three surveys, we find that the mean OE WTP consistently declines from the original survey to the first retest (−7 %) over

[3] The outcome of the Mann-Whitney test statistic is 2.143 ($p = 0.032$).

**Table 8** Mean WTP values and their 95% confidence intervals in CHF/household/year in the test and retests for the policy scenario of low environmental risk in the whole country based on the estimated DCE choice models in WTP space and the OE WTP elicitation format

| Sample test | $N_t$ | WTP (95 % CI) | Sample retest | $N_r$ | WTP (95 % CI) | Wald test | |
|---|---|---|---|---|---|---|---|
| | | | | | | Test statistic | Prob > $\lvert \chi^2 \rvert$ |
| *DCE based WTP* | | | | | | | |
| Original survey | 304 | 105.6 (86.2–125.0) | First retest | 304 | 140.9 (118.4–163.5) | 6.07 | 0.001 |
| Original survey | 245 | 106.9 (80.7–133.1) | Second retest | 245 | 86.7 (72.0–101.4) | 1.82 | 0.068 |
| Original survey | 131 | 102.7 (72.1–133.2) | First retest | 131 | 76.7 (49.9–103.5) | 3.48 | 0.001 |
| Original survey | 131 | 102.7 (72.1–133.2) | Second retest | 131 | 86.8 (59.9–113.6) | 2.13 | 0.033 |
| First retest | 131 | 76.7 (49.9–103.5) | Second retest | 131 | 86.8 (59.9–113.6) | −20.25 | 0.001 |
| *OE WTP* | | | | | | | |
| Original survey | 304 | 113.3 (99.5–127.1) | First retest | 304 | 99.6 (89.2–110.0) | 3.44 | 0.064 |
| Original survey | 137 | 103.4 (81.5–125.4) | Second retest | 137 | 94.7 (76.9–112.6) | 0.40 | 0.526 |
| Original survey | 69 | 103.0 (72.1–134.0) | First retest | 69 | 95.9 (75.8–116.0) | 0.20 | 0.654 |
| Original survey | 69 | 103.0 (72.1–134.0) | Second retest | 69 | 84.1 (68.9–99.3) | 1.38 | 0.240 |
| First retest | 69 | 95.9 (75.8–116.0) | Second retest | 69 | 84.1 (68.9–99.3) | 1.79 | 0.181 |

a time period of 6 months and from the first to the second retest (−12 %) over 18 months. Moreover, for those respondents who participated in all three surveys, mean OE and DCE based WTP are remarkably similar in the original survey and the second retest two years later.[4] The main difference between the OE and DCE based WTP is that the drop in OE WTP after 6 months is substantially less (7 instead of 25 %) and continues to drop further by 12 % in the second retest, resulting in a decline of 18 % over a time period of two years. Contrary to the DCE results, the differences between the mean OE WTP values across the three surveys are not statistically significant at the 5 % level based on the Wald test. Hence, conform previous findings in the test-retest SP literature, the mean OE WTP values decrease, but are statistically speaking the same over a time period of two years based on the sample of respondents who participated in all three surveys. This confirms the third hypothesis for the OE WTP elicitation format that these WTP values are stable over 6, 18 and 24 months.

## 5 Discussion and Conclusions

This paper tests the reliability of the assumption commonly made in the value transfer literature that individuals' preferences are consistent and stable over time. New in this study is that we test preference stability based on two commonly applied elicitation formats at three different points in time. Exactly 6 and 24 months after conducting the original SP survey, using both a DCE and OE WTP question, the same sample of respondents was asked to participate in the same survey again. This allowed us to test the temporal stability of SP and WTP values for the same environmental change in the short and slightly longer term. In addition, we test possible self-selection bias, choice consistency between and within surveys, preference and scale stability, and ordering effects by including the OE WTP question both before and after the DCE.

Similar to the anchoring effects found in the early SP literature where an OE WTP followed a sequence of DC WTP questions (e.g. Green et al. 1998), the stated OE WTP values are significantly anchored on the price levels of the preferred alternative in the last choice set of the DCE. Moreover, there is evidence of significant procedural bias: asking the OE WTP question before the DCE results in a significantly higher mean WTP than asking the same OE WTP question after the DCE. Only a few studies exist that test for such ordering effects and their results are mixed. For example, Hynes et al. (2011) rotated the order of the DCE and OE WTP question in their survey and found no significant difference in WTP estimates, suggesting that no ordering effects are present. On the other hand, Metcalfe et al. (2012) included a payment card and a DC WTP question before and after a DCE and found significantly different WTP values across their split samples. An important caveat of our study is that this ordering effect was only tested in the third survey, making it hard to draw conclusions about the temporal reliability of the two WTP elicitation formats independently.

Our results show that 6 months after the original survey 63 % of the choices are identical and hence strictly consistent, which reduces to 59 % after 24 months. This is comparable to findings in previous DCE studies in the environmental valuation literature (Liebe et al. 2012 and Schaafsma et al. 2014). New in this study is that besides testing choice consistency across the three surveys, we also test choice consistency within the same survey. This yields a substantially higher consistency rate of 80 %, which is somewhat higher than the 73 % reported

[4] The number of observations is substantially lower here than for the DCE findings due to the fact that almost half of the second retest sample who answered the OE WTP question before the DCE was omitted from the analysis presented here.

in Brouwer et al. (2010). Compared to Brown et al. (2008), who report an inconsistency rate of 15 % at the start of their DCE which is halved after 30 choice repetitions due to preference learning, a considerable share of choice inconsistency in our study hence seems to occur within the time frame of just a few minutes.

The study confirms prior expectations that a higher degree of choice certainty and bigger differences between choice alternatives and hence lower choice complexity yield more consistent choices. However, the latter effect fades away in time. The link between choice complexity and choice (in)consistency has been tested in several DCE studies (e.g., DeShazo and Fermo 2002; Louviere 2008; Dellaert et al. 2012. Most of these studies focus more broadly on design complexity dimensions, such as the number of choice sets, attributes and attribute levels, and their impact on the estimated model's error variance. The results so far are mixed (e.g. Swait and Adamowicz 2001; Dellaert et al. 2012). Dellaert et al. (1999) find, for example, that choice consistency decreases as bid level differences increase and absolute bid levels increase. This outcome is different from similar tests, which invariably show that larger differences between alternatives result in higher choice certainty (e.g. Lundhede et al. 2008; Brouwer et al. 2010).

Despite a considerable share of inconsistent choices between the three surveys, the underlying preference parameters in the estimated random utility models appear to be stable over a time period of 18 and 24 months. This is likely due to the strict definition of choice consistency, requiring identical choices in the test and retest. The null hypothesis of equal scale parameters cannot be rejected either at the 5 % significance level over the same time period. The mean WTP values derived from the DCE across the three surveys are nevertheless significantly different and show a clear decline over the two-year time period. This has important implications for the practical use of DCE based welfare measures in benefits transfer. Although DCEs have been argued to be more suitable for benefits transfer than direct WTP elicitation formats since they allow for the modification of context-specific policy characteristics in the utility functions (e.g. Hanley et al. 2001), they seem to have a shorter expiry date. Like the DCE based WTP values, the OE WTP values also decrease in time, but more steadily and not in a significant way based on the same testing procedure. Hence, although the overall decline in estimated OE and DCE WTP values over the 2-year time period is similar, the OE WTP values remain stable.

Finally, directions for future research include more test-retest studies in which surveys are conducted among the same sample of respondents at multiple points and over longer periods of time. Repeating surveys more than once provides better insight into preference dynamics over time and helps to determine the time frame during which benefits transfer based on specific WTP elicitation formats remains valid. This clearly is an underexposed area of research with direct relevance for the practice of benefits transfer. Although our study suggests that the DCE has limited use for benefits transfer over a longer period of time than tested so far in the DCE literature, more empirical evidence on the topic is needed to come to a final conclusion, especially in view of the fact that the estimated WTP values and their evolution over time based on the two elicitation formats are in the same order of magnitude. Ideally, the stability of welfare measures is also tested independently based on different elicitation formats over longer periods of time so as to evaluate their suitability for inclusion in benefits transfer.

**Acknowledgments** This study was funded by the Swiss Federal Institute of Aquatic Science and Technology (Eawag) in Zürich. We are grateful to two anonymous reviewers and the co-editor Christian Vossler for their thorough reading and feedback on previous versions of this paper.

## References

Bateman IJ, Langford IH, Turner RK, Willis KG, Garrod GD (1995) Elicitation and truncation effects in contingent valuation studies. Ecol Econ 12(2):161–179

Berrens RP, Bohara AK, Silva CL, Brookshire D, McKee M (2000) Contingent values for New Mexico instream flows: with tests of scope, group-size reminder and temporal reliability. J Environ Manag 58(1):73–90

Bhat CR (2001) Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. Transp Res B 35(7):677–695

Bliem M, Getzner M, Rodiga-Laßnig P (2012) Temporal stability of individual preferences for river restoration in Austria using a choice experiment. J Environ Manag 103:65–73

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. Ecol Econ 60(2):399–406

Brouwer R (2012) Constructed preference stability: a test-retest. J Environ Econ Policy 1(1):70–84

Brouwer R, Bateman IJ (2005) The temporal stability and transferability of models of willingness to pay for flood control and wetland conservation. Water Resour Res doi:10.1029/2004WR003466

Brouwer R, Dekker T, Rolfe J, Windle J (2010) Choice certainty and consistency in repeated choice experiments. Environ Resour Econ 46(1):93–109

Brouwer R, Job F, van der Kroon B, Johnston R (2015) Comparing willingness to pay for improved drinking water quality using stated preference methods in rural and urban Kenya. Appl Health Econ Health Policy 13(1):81–94

Brown TC, Kingsley D, Peterson GL, Flores NE, Clarke A, Birjulin A (2008) Reliability of individual valuations of public and private goods: choice consistency, response time, and preference refinement. J Public Econ 92(7):1595–1606

Daly AJ, Hess S, Train KE (2012) Assuring finite moments for willingness to pay in random coefficients models. Transportation 39(1):19–31

Dekker T, Hess S, Brouwer R, Hofkes M (2016) Decision uncertainty in multi-attribute stated preference studies. Resour Energy Econ 43:57–73

Dellaert BGC, Brazell JD, Louviere JJ (1999) The effect of attribute variation on consumer choice consistency. Mark Lett 10(2):139–147

Dellaert BG, Donkers B, van Soest A (2012) Complexity effects in choice experiment-based models. J Mark Res 49(3):424–434

DeShazo JR, Fermo G (2002) Designing choice sets for stated preference methods: the effects of complexity on choice consistency. J Environ Econ Manag 44(1):123–143

Fetene GM, Olsen SB, Bonnichsen O (2014) Disentangling the pure time effect from site and preference heterogeneity effects in benefit transfer: an empirical investigation of transferability. Environ Resour Econ 59(4):583–611

Foster V, Mourato S (2003) Elicitation format and sensitivity to scope. Environ Resour Econ 24(2):141–160

Green D, Jacowitz KE, Kahneman D, McFadden D (1998) Referendum contingent valuation, anchoring, and willingness to pay for public goods. Resour Energy Econ 20:85–116

Hanley N, Mourato S, Wright RE (2001) Choice modelling approaches: A superior alternative for environmental valuation? J Econ Surv 15(3):435–462

Hensher DA, Rose JM, Greene WH (2005) Applied choice analysis: a primer. Cambridge University Press, Cambridge

Hoeffler S, Ariely D (1999) Constructing stable preferences: a look into dimensions of experience and their impact on preference stability. J Consum Psychol 8(2):113–139

Hollender J, Zimmermann SG, Koepke S, Krauss M, McArdell CS, Ort C, Singer H, von Gunten U, Siegrist H (2009) Elimination of organic micropollutants in a municipal wastewater treatment plant upgraded with a full-scale post-ozonation followed by sand filtration. Environ Sci Technol 43(20):7862–7869

Holmes TP, Boyle KJ (2005) Dynamic learning and context-dependence in sequential, attribute-based, stated-preference valuation questions. Land Econ 81(1):114–126

Hynes S, Campbell D, Howley P (2011) A holistic versus an attribute-based approach to agri-environmental policy valuation: Do welfare estimates differ? J Agric Econ 62(2):305–329

Kidd KA, Blanchfield PJ, Mills KH, Palace VP, Evans RE, Lazorchak JM, Flick RW (2007) Collapse of a fish population after exposure to a synthetic estrogen. Proc Natl Acad Sci USA 104(21):8897–8901

Liebe U, Meyerhoff J, Hartje V (2012) Test–retest reliability of choice experiments in environmental valuation. Environ Resour Econ 53(3):389–407

List JA (2003) Does market experience eliminate market anomalies? Q J Econ 118:41–71

Logar I, Brouwer R, Maurer M, Ort C (2014) Cost-benefit analysis of the Swiss national policy on reducing micropollutants in treated wastewater. Environ Sci Technol 48(21):12500–12508

Louviere JJ, Hensher DA, Swait JD (2000) Stated choice methods: analysis and application. Cambridge University Press, Cambridge

Louviere JJ et al (2008) Designing discrete choice experiments: Do optimal designs come at a price? J Consum Res 35(2):360–375

Lundhede TH, Olsen SB, Jacobsen JB, Thorsen J (2008) Choice experiments and certainty in choice: a test of the influence of utility difference on self-reported certainty levels and evaluations of three recoding approaches to handle uncertrain responses. Paper presented at the 16th annual conference of the European association of environmental and resource economists (EAERE), June 25–28 2008, Gothenburg, Sweden

McConnell KE, Strand IE, Valdes S (1998) Testing temporal reliability and carryover effect: the role of correlated responses in test–retest reliability studies. Environ Resour Econ 12(3):357–374

McFadden D (1974) Conditional logit analysis of qualitative choice behavior. In: Zarembka P (ed) Frontiers in econometrics. Academic Press, New York, pp 105–142

McFadden D, Train KE (2000) Mixed MNL models for discrete response. J Appl Econ 15(5):447–470

Messonnier ML, Bergstrom JC, Cornwell CM, Teasley RJ, Cordell HK (2000) Survey response-related biases in contingent valuation: concepts, remedies, and empirical application to valuing aquatic plant management. Am J Agric Econ 83:438–450

Metcalfe P, Baker W, Andrews K, Atkinson G, Bateman I, Butler S, Carson R, East J, Gueron Y, Sheldon R, Train K (2012) An assessment of the nonmarket benefits of the water framework directive for households in England and Wales. Water Resour Res 48(3):W03526. doi:10.1029/2010WR009592

Mogas J, Riera P, Bennett J (2006) A comparison of contingent valuation and choice modelling with second-order interactions. J Forest Econ 12(1):5–30

Mørbak MR, Olsen SB (2014) A within sample investigation of test–retest reliability in choice experiment surveys with real economic incentives. Aust J Agric Resour Econ 56:1–18

Schaafsma M, Brouwer R, Liekens I, Denocker L (2014) Temporal stability of preferences and willingness to pay for natural areas in choice experiments: a test–retest. Resour Energy Econ 38:243–260

Swait J, Adamowicz W (2001) The influence of task complexity on consumer choice: a latent class model of decision strategy switching. J Consum Res 28(1):135–148

Swait J, Louviere J (1993) The role of the scale parameter in the estimation and comparison of multinomial logit models. J Mark Res 30(3):305–314

Whitehead JC, Hoban TJ (1999) Testing for temporal reliability in contingent valuation with time for changes in factors affecting demand. Land Econ 75(3):453–465