Environ Resource Econ (2012) 53:389–407
DOI 10.1007/s10640-012-9567-1

Test–Retest Reliability of Choice Experiments in Environmental Valuation

Ulf Liebe · Jürgen Meyerhoff · Volkmar Hartje

Accepted: 21 May 2012 / Published online: 6 June 2012

Abstract The paper presents the results of the first test–retest study on choice experiments in environmental valuation. In a survey concerning landscape externalities of onshore wind power in central Germany, respondents answered the same five choice sets at two different points in time. Each choice set comprised three alternatives described by five attributes, and the time interval between the test and the retest was eleven months. The analysis takes place at three different levels, investigating choice consistency at the choice task level and repeatability of the latent construct utility at the level of parametric models as well as at the level of willingness-to-pay estimates. At the choice task level we observed 59 % identical choices. The parametric analysis shows that the test and retest estimates are not equal, even when we control for scale, that is, differences in the error variance. However, comparing the marginal willingness-to-pay estimates among test and retest reveals only a statistically significant difference for one of the attributes. Overall, this indicates a moderate test–retest reliability taking into account that consistency at the choice task level overlooks the stochastic nature of the process underlying discrete choice experiments.

Keywords Choice experiment · Environmental valuation · Test–retest reliability · Wind power

JEL Classification C8 · Q0 · Q5

U. Liebe (✉)
Department of Agricultural Economics and Rural Development, Georg-August-Universität Göttingen, Platz der Göttinger Sieben 5, 37073 Göttingen, Germany
e-mail: uliebe@uni-goettingen.de

U. Liebe
Department of Rural Sociology, Universität Kassel, Witzenhausen, Germany

J. Meyerhoff · V. Hartje
Institute for Landscape and Environmental Planning, Technische Universität Berlin, Berlin, Germany

1 Introduction

Choice experiments, used as an alternative to the contingent valuation method in environmental valuation, are undergoing comprehensive scrutiny. Aspects currently investigated include for example the effect of the experimental design (e.g. Ferrini and Scarpa 2007; Lusk and Norwood 2005), the choice task complexity (e.g. DeShazo and Fermo 2002; Boxall et al. 2009) or the non-attendance to choice attributes (e.g. Hensher et al. 2005; Campbell et al. 2008) on welfare estimates. These studies mainly examine the validity of the method, that is, the relation between the measure and the underlying construct. In contrast, test–retest reliability has to the best of our knowledge not been investigated in any choice experiment study in environmental valuation.

That the test–retest reliability of choice experiments has so far not been investigated is surprising considering that reliability is one of the classical and most fundamental problems in empirical research. As Guttman (1945, 256) puts it, "[i]n dealing with empirical data in any field, the question should be raised: if the experiment were to be repeated, how much variation would there be in the results?" Furthermore, as stated preference studies in general merely provide information about preferences at a particular point of time, no information is available about the temporal stability of the elicited estimates. Moreover, it is important for decision makers and resource managers to know whether people value the same environmental changes similarly over time.

In the present study test–retest reliability of choice experiment responses is evaluated at different levels. In line with three test–retest studies from health economics (Bryan et al. 2000; Ryan et al. 2006; Skjoldborg et al. 2009, see Sect. 2.3) we first analyse choice consistency at the choice-set level. Here the question is how many identical choices would be observed. Second, at the level of parametric models we investigate similarity of utility functions. Here the question is whether the repeated choice experiment would result in corresponding logit models that give statistically not significantly different coefficients for utility parameters. Third, at the level of willingness to pay we investigate whether the repeated choice experiment would give statistically not significantly different estimates. However, using these three criteria for analysing test–retest reliability one has to keep in mind that investigating choice consistency at the choice task level overlooks the stochastic nature of the choice process as it is assumed by discrete choice models based on random utility theory. Thus, this criterion is likely to evaluate reliability too critically.

To some extent test–retest reliability has been investigated with respect to the contingent valuation method (McConnell et al. 1998; Jorgensen et al. 2004; Brouwer and Bateman 2005). Results show a wide range from fair to very good reliability. Studies dealing with test–retest reliability of responses to choice experiments have so far all been carried out in the area of health economics (e.g. Bryan et al. 2000; Ryan et al. 2006; Skjoldborg et al. 2009). Within environmental valuation only one study, dealing with preferences for water quality, has analysed temporal reliability across two independently drawn samples (Bliem et al. 2012). However, as the authors drew two independent samples the survey was not presented to the same respondents in the retest. The study therefore indicates whether the overall preferences in the population remain stable but it does not indicate whether or not the same respondents would answer identically at both points in time. In addition to Bliem et al. (2012) a couple of studies have investigated the stability of choices within the same choice experiment by repeating a choice set at the end of the sequence of choice sets faced by each respondent (e.g. Scarpa et al. 2007a). While the same respondent faces the same choice set again in these studies, they are not investigating test–retest reliability due to the short time interval between the test and the retest. Similarly,

Carlsson et al. (2012) present the same respondents the identical block of eight choice sets twice within the same survey. As the two blocks followed in succession, this study does not qualify as a test–retest study either. The present paper therefore is to our knowledge the first study that presented the same respondent the same choice sets with a significant time interval between test and retest, thereby decreasing the probability of carry-over effects.

As has been mentioned above an analysis of whether the valuation differs between test and retest is carried out at various levels. Firstly, we compare the responses to the test and the retest at the choice set level. Secondly, the parameter estimates from the test and retest are compared based on separate models and on a pooled model. Thirdly, we calculate the marginal willingness-to-pay values for the attribute levels and compare them between test and retest. The remaining paper is organized as follows: Sect. 2 introduces the method of test–retest reliability in general and with respect to choice experiments in particular. It also briefly presents findings from previous contingent valuation studies in environmental economics and from three choice experiments in the field of health economics. While Sect. 3 describes the choice experiment concerning landscape externalities from onshore wind power, Sect. 4 presents the results of the test–retest analysis. The paper concludes in Sect. 5.

2 Test–Retest Method and Non-Market Valuation

2.1 Test–Retest Method

In order to evaluate reliability using the test–retest method, the same survey instrument is given to the same group of respondents at (at least) two points in time where the first point represents the test and the second point the retest (e.g. Yu 2005 for an overview). Measuring the correlation between the results of the test and the retest, researchers generally depend on a reliability coefficient that represents "the ratio of the variance of the true scores on a test to the variance of observed scores: $r = \sigma^2_r / \sigma^2_x$, where r is the theoretical reliability of the test, σ^2_r the variance of the true scores and σ^2_x the variance of the observed scores" (Kaplan and Saccuzzo 2008, 108). The classification of test–retest reliability varies typically between 'poor' test–retest reliability, that is, no agreement between test and retest results, and 'perfect' test–retest reliability, that is, a perfect agreement of test and retest results. Gradations in between are arbitrary. Landis and Koch (1977, 165), for instance, provide the following benchmarks for classifying the strength of agreement based on the Kappa statistic: 0.00–0.20 'slight', 0.21–0.40 'fair', 0.41–0.60 'moderate', 0.61–0.80 'substantial', 0.81–1.00 'almost perfect' agreement.

However, test–retest results can be biased (e.g. Kaplan and Saccuzzo 2008, 110; Ary et al. 2009, 242). If respondents remember their answers from the first test in the retest a carry-over or memory effect can occur. This would typically lead to an overestimation of reliability, that is, a higher correlation between results of the test and retest. In order to avoid a memory effect, the test and the retest should not take place within a time span that is too short. If the time interval, on the other hand, increases too much, external events such as changing economic conditions or changes of respondents' characteristics such as attitudes might affect findings. Researchers therefore have to take into account events and factors that might have changed between the test and the retest when interpreting the results of test–retest studies.

2.2 Test–Retest Reliability and Choice Experiments

Test–retest studies investigating the reliability of contingent valuation responses have employed various tests. In their overview McConnell et al. (1998) name, among others, the correlation between responses to contingent valuation questions, equality of parameters between the distinct samples, and the equality of willingness-to-pay estimates between the samples as tests. Similarly, test–retest studies regarding the results of choice experiments can test the reliability at various levels as it has been done in several studies from health economics (Bryan et al. 2000; Ryan et al. 2006; Skjoldborg et al. 2009, see Sect. 2.3). Firstly, the correlation between test and retest choices can be measured at the level of choice sets. To what extent do we observe identical choices between the test and retest for single choice sets and across all choice sets of a choice experiment? Symmetry between the choices made at the test and the retest would indicate a high level of reliability and preference stability at the level of choice sets. Secondly, at the level of individuals it can be analyzed to what extent respondents answered each choice set identically in the test and retest. The highest level of test–retest reliability would be obtained when each individual answers the choice sets in the test and retest exactly in the same way. Thirdly, it can be tested whether test and retest results differ at the level of parameter estimation and willingness-to-pay estimates. High test–retest reliability is given if coefficient estimates and willingness-to-pay values do not significantly differ between the test and the retest. Taking into account differences in scale, that is the variance of the error term, between test and retest can indicate whether changes in the parameter estimates are a result of preference changes or of a lower error variance because of learning effects, for instance. At all three levels responses to debriefing questions—regarding, for example, attitudes toward the good in question or events that occurred between the test and the retest—can be taken into account in order to determine whether changes in attitudes or external events affect test–retest reliability.

Apart from these similarities between test–retest studies of contingent valuation and choice experiments some differences must be highlighted. Generally, in a typical choice experiment respondents face a sequence of choices depending on the experimental design. The majority of studies presents respondents four to eight choice sets. In contrast, typical contingent valuation studies present respondents one or two questions to elicit their willingness to pay. The likelihood that people do not give exactly the same responses at both points in time may therefore be higher in a choice experiment. Moreover, respondents generally have the opportunity to choose between at least two alternatives when facing a choice experiment. Depending on the design of the experiment, a typical choice set may have two to five alternatives (in most cases including an option not to choose). Respondents may therefore choose, for example, a non-status quo alternative in the same choice set in both the test and the retest. However, the chosen non-status quo alternatives can differ. Respondents would still be willing to pay for an environmental change but their preferences with respect to the attribute levels might have changed slightly. If reliability would be tested only at the level of equal choices this would overlook the stochastic nature of the process underlying discrete choice experiments and is therefore likely to underestimate reliability. Studies applying joint error components for the hypothetical alternatives in the econometric analysis, for example, have provided ample evidence that the utilities of these alternatives are often highly correlated (e.g. Scarpa et al. 2007b). Choosing a different non-status quo alternative therefore indicates a different choice but can still express similar benefits derived from the environmental change investigated. Requiring a perfect agreement of choices at the individual level, that is, each respondent makes identical choices across all choice sets in the test and retest, might therefore be too restrictive. Changes of choices at the individual level might aggregate to non-significant differences of

parameter estimates and willingness-to-pay values for attributes. The most informative way to investigate test–retest reliability of choice experiments is a cumulative strategy comprising the level of single choice sets and blocks resulting from the experimental design, the level of choices per individual, and the level of parameter and willingness-to-pay estimates.

2.3 Previous Findings in Non-Market Valuation

Several studies examine temporal reliability with regard to contingent valuation. For this purpose the same survey is carried out at least at two points in time using independent samples. McConnell et al. (1998) review temporal reliability studies in general and test–retest reliability studies in particular. They point to several studies with positive evidence for temporal reliability based on at least two independent samples taken at least at two different points in time. In contrast, a study by Brouwer and Bateman (2005) comparing results across a five-year period, a very long time span for reliability studies, demonstrates that willingness-to-pay estimates can significantly change over time. As an extension of the study by McConnell et al. (1998), Jorgensen et al. (2004, p. 43) reviewed eight test–retest studies with each study comprising one to five test–retests. The reported reliability coefficients range from 0.30 to 0.95. The investigated studies value quasi-private goods such as hunting permits as well as public goods such as air quality. Different question formats such as open-ended or dichotomous questions to obtain respondents' willingness to pay are applied, and the time interval between the test and the retest is a minimum of two weeks and a maximum of three years. Overall, the review by Jorgensen et al. (2004) indicates no clear pattern with respect to the correlation between study characteristics and test–retest reliability when contingent valuation is employed.

Bliem et al. (2012) are so far the only authors who have investigated temporal stability of choice experiment results in the context of environmental valuation. They used two independent samples at two different points in time. The time lag between the two identical web-based surveys was one year. The sample size in 2007 was 506 cases and 410 cases in 2008. The authors do not find remarkable differences between the significance of the choice attributes or between willingness-to-pay estimates in both surveys. The only studies that investigated test–retest reliability of choice experiments by presenting the same choice sets to the same individuals are all from health economics. They are briefly summarised in the following paragraphs.

(1) Bryan et al. (2000) investigate the test–retest reliability of a discrete choice experiment in health care, more precisely, 'preferences for treatment options for patients with knee injuries'. The choice experiment was conducted in the UK with undergraduate students (N = 585) who were studying to become sport scientists or teachers and included four attributes, two options per choice set, and three questionnaires (i.e. three survey waves). In the first questionnaire, respondents answered eight choice sets. Next, they were asked to answer another eight sets in a second questionnaire immediately after finishing the first one. In the second questionnaire, four of the eight choice sets were exact duplicates of the first sets. This very short time interval between the first and second questionnaire raises the likelihood that a carry-over effect occurs. It follows that a comparison between the first and second questionnaire in the study by Bryan et al. can be considered a poor measure of test–retest reliability (similar to studies on stability of choices within one and the same experiment, e.g. Scarpa et al. (2007a). However, the majority of respondents also answered a third questionnaire two weeks after the first two questionnaires. This time the questionnaire included 12 choice sets

exactly duplicating the choice sets from the first and second questionnaires. The results show that 57 % of the respondents chose exactly the same alternatives in the first and second questionnaire. In the third survey wave, 22 % did so in all 12 choice tasks and 32 % in 11 out of 12 choice tasks. In general, there were 86 % congruent responses between the first and second questionnaire as well as the first/second and the third questionnaire. The reliability coefficients are 0.71 and 0.65, respectively. This can be classified as substantial reliability. Finally, there are no notable differences between estimates derived from multivariate models based on the data from the three survey waves.

(2) Ryan et al. (2006) conducted a discrete choice experiment to value the quality of outcomes regarding social services for elderly people in the UK. The choice experiment consisted of 14 choice sets which were split equally between two questionnaires, that is, seven sets per questionnaire. Each choice set had two alternatives with five attributes and each attribute had three levels. The sample comprised 357 people aged 60 and older; 47 of these interviewees filled out a second choice experiment between 11 and 60 days after the first interview date. The findings indicate substantial test–retest reliability. Out of 375 duplicated choices made by the 47 interviewees, 82 % were in exact agreement; this gives a reliability coefficient of 0.64. Furthermore, the results of regressions models based on the test and retest were not significantly different from each other.

(3) Skjoldborg et al. (2009) investigate test–retest reliability of a choice experiment concerning arthritis medication in Odense, Denmark. Each choice set included two alternatives with six attributes, with the number of attribute levels ranging from two to 18 levels. The 62 choice sets obtained from the experimental design were blocked in eight groups with ten choice sets in each group. The choice tasks were answered by 178 patients diagnosed with rheumatoid arthritis. Respondents participated in three interviews. The time period between the test studies and retest studies was four months; 145 (81 %) and 130 (73 %) of the 178 respondents in the first interview participated in the second and third interview, respectively. Skjoldborg et al. (2009) report 76 % congruent choices in the second interview and 87 % in the third interview. The regression results and willingness-to-pay estimates did not significantly differ between interviews one, two, and three.

The three studies in health economics indicate substantial test–retest reliability of the applied choice experiments. However, all three studies differ considerably with respect to survey design and settings. The time span between the test and retest varies from conducting the retest immediately following the test to conducting it four months after the test. While results might be prone to carry-over effects especially in the case of the shortest time span, even a four months time span might be rather short to test for test–retest reliability. Further, all three choice experiments included just two options per choice set. This might work in favour of positive test–retest reliability. In contrast, most applications in environmental valuation present three alternatives on a choice set.

Another aspect might limit comparability with environmental valuation studies. The studies were conducted in the field of health economics, thus the benefits arising from the goods under consideration are based on use values. In contrast, benefits of environmental goods often include both use and non-use values. Moreover, they typically comprise non-market goods, compared to market goods that are often the subject of studies in health economics. Thus, the question emerges as to whether test–retest results differ with respect to the type of good under consideration.

3 Study Design and Data

The data were collected in a choice experiment carried out to determine landscape externalities from onshore wind power in the region of Nordhessen, located in central Germany (see Meyerhoff et al. 2010 for a comprehensive description of the study).

Table 1 gives an overview of all attributes as well as their levels (see Table 8 in the appendix for an example of a choice set used in the survey). Programme A describes for each choice set how wind power would develop until 2020 in the study region if respondents did not decide otherwise. Respondents were informed that Programme A would allow electricity to be produced from wind power at low costs and hence choosing this alternative requires no surcharges. It follows that Programme A can be interpreted as a kind of status quo or no-cost alternative. Whereas Programme A always has the same attribute levels, Programmes B and C restrict wind power generation at least with respect to one attribute compared to Programme

Table 1 Attributes and their levels used in the choice experiment

Attributes	Information given in survey	Levels
Size of wind farms	Larger wind farms generally lower the costs of electricity production but the bigger they are, the greater their potential influence on the landscape; when farms are larger in total, fewer farms are needed to produce the same amount of electricity in Nordhessen	*Large farms (16–18 mills)* Medium farms (10–12 mills) Small farms (4–6 mills)
Maximum height of turbines	The higher turbines are, the more electricity can be generated as winds are stronger and more constant at higher altitudes. Thus, fewer turbines are needed to produce a certain amount of electricity. On the other hand, visibility increases with height	110 m 150 m *200* m
Impact on red kite population	Turbines would not be installed in conservation areas but even outside these areas conflicts may arise. For example, negative impacts on the red kite, a predatory bird with one of its main habitats in the region, may lead to a decreasing population	5 % *10 %* 15 % Reduction of red kite population
Minimum distance from residential areas	Regulations stipulate that turbines have to be at a minimum distance from residential areas in order to avoid adverse effects, for example through noise or shading. Programme A with a minimum distance of 750 m complies with these regulations. Visibility would diminish with greater distances	*750* m 1.100 m 1.500 m
Monthly surcharge to power bill beginning in 2009	Programme A presents today's state of technology and thus allows electricity to be generated efficiently. Programmes B and C would lead to higher costs, for example, for infrastructures such as longer power cables, and thus require a surcharge to the monthly power bill	*€0* €1 €2.5 €4 €6

Levels in italics are those of Programme A, the information given in the survey is presented here in a summarised form

A. Besides a price attribute, the alternatives were described by four attributes characterising different impacts of wind power: size of wind farms, maximum height of turbine, impact on the red kite population, and the minimum distance from residential areas.

The attributes and their levels were selected on the basis of findings of three focus groups with a total of 25 participants as well as scientific information (e.g. information about the impact of turbines on birds). Respondents were informed that all three programmes on a choice set would result in the same amount of avoided carbon dioxide per year (550,000 tons). The implementation of Programme B or C would require a monthly surcharge to their power bill as both alternatives imply rising costs for electricity production, for example due to higher spending on longer power cables when turbines are built further away from residential areas. Designing the choice sets we followed a procedure outlined in Johnson et al. (2007). It uses a fractional factorial design and seeks to efficiently allocate alternatives to choice sets assuming that the levels are evenly spaced levels as well as assuming zero valued parameter priors. The final design consisted of 40 choice sets that were blocked into eight subgroups with five choice sets each. In the main survey, the test, each of these blocks was presented to at least 44 respondents. However, the design applied in the test–retest study is only a part of the original design as not all respondents of the main survey participated in the retest and thus not all blocks are used equally frequently.

The main survey, the test, was conducted through telephone interviews in May and June 2008 by a survey organisation. The target population was defined as adults aged 18 or over living in Nordhessen, the study area. Households were contacted by random digit dialling, and the target person was selected via the Kish selection grid method. If the person agreed a date for the main interview was arranged. Information about the objective of the survey, detailed descriptions of the attributes and the choice sets were mailed to the target person. The interview was structured as follows: Respondents were first presented the choice sets and subsequently a few questions concerning, among other things, their experience with and attitudes towards wind power. Finally, socio-demographics were requested.

Overall, 355 interviews were conducted (corresponding to a response rate of 35 %). At the end of the interview all participants were asked whether they would agree to be contacted again by the survey company; 95 % responded positively. The retest was carried out in April and May 2009, eleven months after the main survey. The same survey organisation was asked to conduct around 170 interviews with respondents from the first survey. This number of interviews resulted from the budget constraints of the retest study. Overall, 298 respondents had to be contacted to carry out 172 retest interviews. Among those unsuccessfully contacted, 41 % could not be reached and 59 % refused to give another interview.

The procedure in the retest survey was exactly the same as in the main survey: Respondents were mailed information about wind power generation in the study region, detailed descriptions of the attributes and the choice sets. The telephone interview began with the choice sets and each participant valued exactly the same sets as in the main survey. Next, in addition to socio-demographics requested to check whether the same person is responding to the retest, the questionnaire included several debriefing questions. These questions aimed at any concern with wind power between the test and the retest, changes in the personal financial situation of respondents, and the perceived impact of the financial crisis—which occurred between the time points of the test and retest—on the respondent's personal situation. This information can help to determine to what extent changes in attitudes toward wind power, changed budget constraints or external effects affect test–retest results.

Table 2 provides descriptive statistics on the subgroups of respondents who participated in both the test and the retest and those who only participated in the test. In order to test

Table 2 Group comparison of participants and non-participants in the retest

	Participants N = 172		Non-Participants N = 183		Sig. diff.
	Mean	SD	Mean	SD	
Gender (1 = female)	0.55	0.50	0.44	0.50	Yes
Age in years	50.00	14.67	46.52	17.00	Yes
Education (1 = secondary school+)	0.51	0.50	0.46	0.50	No
People per household	2.58	1.24	2.64	1.32	No
Income (missing values imputed)	2523.04	981.87	2309.75	973.07	Yes
Lives in a city (1 = yes)	0.16	0.36	0.20	0.40	No
Lives near wind turbines (1 = yes)	0.22	0.42	0.17	0.38	No
Saw wind turb. in last 4 weeks (1 = yes)	0.85	0.35	0.86	0.34	No
Donation for env. last year (1 = yes)	0.38	0.49	0.30	0.46	No
Env. group member (1 = yes)	0.15	0.35	0.09	0.29	No
Dealing with topic since the test (1 = yes)	0.58	0.50			
Change of financial situation (1 = yes)	0.25	0.43			
Affected by financial crisis (1 = yes)	0.17	0.38			

Due to a missing value, the group of participants comprises 171 respondents for the variable "affected by financial crisis." All variables are binary coded except age (min = 19, max = 82 for participants; min = 18, max = 83 for non-participants), income (min = 500, max = 4, 500) and people per household (min = 1, max = 7 for participants; min = 1, max = 6 for non-participants). "yes" in the column "Sig. diff." means that differences are significant at least at the 5 % level. A chi-square test was applied for the binary coded variables and mean comparison tests and Wilcoxon–Mann–Whitney tests for age, income, and people per household

for differences between the subgroups at a bivariate level a chi-square test was applied for the binary coded variables and mean comparison tests as well as Wilcoxon–Mann–Whitney tests for the variables age, income, and people per household. The statistics reveal, on the one hand, that the proportion of females is significantly higher among participants in the retest. The same applies to mean age as well as mean net household income. On the other hand, there are no significant differences with respect to education, household size, exposure to wind power (i.e. living near a wind farm and frequency of seeing a wind farm), or environmental activism (i.e. donation behaviour, albeit significant at the 10 % level, and environmental group membership). Overall, the differences between participants and non-participants can be qualified as rather marginal as particularly the variables indicating experience with wind turbines (the subject of the valuation study) do not indicate significant differences.

At a multivariate level none of the variables shows a significant effect at the 5 %-level in a logit model with retest participation as the dependent variable (results are not presented as the model results are not significant). Thus, there is no indication of selection bias in the retest with regard to respondents' characteristics. Finally, as can be seen in Table 2, 58 % of the participants affirmed that they had dealt with the issue of wind power generation since the first survey, 25 % stated that their personal financial situation had changed since then (i.e. their situation was better or worse), and 17 % stated that they felt personally affected by the financial crisis that took place in autumn 2008 (i.e. they answered that they were personally affected by the financial crisis to some or a large extent). We analyse whether these factors have influenced the response behaviour in the retest.

4 Results

4.1 Test–Retest Reliability at the Level of Choice Sets and the Individual Level

A cross-tabulation of the choices made at the test and the retest is presented in Table 3. Since each respondent could choose between three alternatives per choice task and was presented five choice sets a total of 860 choices were made by 172 respondents. As can be seen from Table 3, the majority of choices in the retest match the choices from the test. However, the figures vary between alternatives. The greatest congruence is observed for alternative A, followed by alternatives B and C. We also find that there is no trend towards choosing alternative A, the status-quo and no-cost alternative, in the retest disproportionately often when either alternative B or C was chosen in the test. This is noteworthy because choosing a no-cost alternative instead of an alternative with a positive price can be interpreted as a market exit. If it was the other way around this would indicate a market entry. Respondents in the present study have rarely changed their decision to be willing to pay or not be willing to pay for constraining wind power generation.

Overall, the test and retest contain 59 % identical choices (i.e. $(203 + 180 + 122)/860 = 0.587$). Taking the benchmarks proposed by Landis and Koch (1977) as a baseline this can be classified as a fair agreement and association between choices (kappa statistic of 0.38). The test proposed by Bowker (1948), suitable for dependent samples, does not allow a rejection of the null hypothesis of symmetry between test and retest at the 5 %-significance level with regard to the total of five choice sets (Chi-square = 3.05, 3 df, $p = 0.384$).

However, as Table 4 shows we find falsifications of symmetry between test choices and retest choices at the level of single choices sets. In our study the choice sets were allocated to eight blocks with each block including five choice sets. It turns out that for three out of 40 choice sets the assumption of symmetry does not hold at a 5 % significance level; another three choice sets show significant differences at the 10 % level. Next, comparing choices per

Table 3 Congruent choices across test and retest

Test	Retest			Total
	A	B	C	
A	203	69	49	321
	63.24	21.50	15.26	100
	67.89	21.77	20.08	37.33
B	51	180	73	304
	16.78	59.21	24.01	100
	17.06	56.78	29.92	35.35
C	45	68	122	235
	19.15	28.94	51.91	100
	15.05	21.45	50.00	27.33
Total	299	317	244	860
	34.77	36.86	28.37	100
	100	100	100	100

Each cell contains the absolute number of choices (first value), row percentages with regard to test choices (second value), and column percentages with regard to retest choices (third value)

Table 4 Bowker's test for symmetry at the level of choice sets

	Block of choice sets							
	1	2	3	4	5	6	7	8
Choice set 1	1.000	0.367	0.063	1.000	0.906	1.000	0.031	0.484
Choice set 2	0.625	0.359	1.000	0.688	0.813	0.775	1.000	0.031
Choice set 3	0.672	0.029	0.063	0.125	0.195	0.813	0.219	0.113
Choice set 4	0.813	0.078	0.781	0.219	1.000	0.508	0.766	0.343
Choice set 5	0.551	0.282	0.150	0.219	0.625	0.051	0.689	0.243
Total obs.	115	110	90	85	100	120	120	120

Obs. means observations. Presented are p values of exact Bowker's symmetry tests for each of the 40 choice sets included in the choice experiment (five choice sets per each of the eight blocks)

Table 5 Congruent choices at the level of individuals

Congruent choices	N	Percent	Percent random
0	10	06	13
1	24	14	33
2	35	20	33
3	36	21	16
4	32	19	4
5	35	20	0.4
Total	172	100	

The table gives the number and proportion of respondents with congruent choices in the test and retest. For example, 21 % of all respondents made the same choices in three out of five choice tasks. The last column gives percentage values based on the assumption of random responses

block, in three out of eight blocks we find at least one choice set per block with significant differences at the 5 % level; in another two choice blocks we observed significant differences at the 10 % level. On the one hand the findings amount to a strong symmetry at the level of single choice sets. If individuals choose another alternative in the retest than in the test, these changes tend to be non-systematic. Our finding indicates, on the other hand, that certain choice sets and some of the eight blocks might be prone to non-symmetry of answers. This in turn could have been caused, for example, by differences in choice-set characteristics such as utility balance as a high utility balance makes choices more volatile (Olsen et al. 2011).

We also analysed the congruence of choices between test and retest for each respondent. How many identical choices can be observed for each respondent? Table 5 reports the number of congruent choices at the individual level.[1] It turns out that 20 % of the respondents make five identical choices and 60 % at least three identical choices; 6 % of the respondents make no identical choices at all. Thus, only for a minority of respondents (20 %) choices match perfectly between the test and the retest. However, taking into account the large number of three and four congruent choices, the majority of their choices match between the test in the

[1] Given the choice set and choice block specific effects presented in Table 4 a reviewer noted that Table 5 might reflect choice set and choice block specific effects rather than differences in individuals. This holds true if choice set und choice block effects and individual effects are highly correlated. However, we find no clear pattern that most individuals that were perfectly consistent answered blocks where the most symmetry was observed and that individuals that were least consistent answered blocks where the least symmetry was observed.

retest (i.e. three, four or five choices) for more than half of the respondents (60 %). Further, we observe a higher number of three, four and five congruent choices than would be expected by chance (see the last column in Table 5); a Kolmogorov–Smirnov test indicates with a test statistic of D equal to 0.391 ($p < 0.001$) that the distribution of actual congruent choices significantly differs from the one expected by random responses. This can be interpreted as supporting evidence for test–retest reliability.

Since there was a considerably long time span between the test and the retest, we checked whether our descriptive and bivariate findings are affected by several potential influencing factors. As described in Sect. 3, the debriefing questions aimed at, among other things, concern with wind power since the test, changes in the respondents' personal financial situation, and perceived effects of the financial crisis on respondents' households (see Table 2). Neither these factors nor socio-demographic variables show any noteworthy significant correlation with the number of congruent choices (based on bivariate and multivariate analyses). Thus, effects of changes in the personal situation or 'external events' are not relevant in our study.

4.2 Parametric Analysis and Willingness-to-Pay Estimates

For the parametric analysis we employ error component logit (ECL) models. The ECL model considers heterogeneity with respect to the alternatives while avoiding the restrictive IIA assumption and accounting for the panel character of the data at the same time (Scarpa et al. 2005; Hensher et al. 2007). Models also taking taste heterogeneity with respect to the attributes into account revealed only low improvements in model performance, that is, standard deviations of the random parameters were mainly not significant. Therefore, the more parsimonious ECL model is preferred. Generally, the utility function for this model can be written as

$$U_{ni} = V_{ni} + E_{ni} + \varepsilon_{ni} \tag{1}$$

where V_{ni} is the systematic component of utility, E_{ni} are the error components, and ε_{ni} is the same type 1 extreme value term as in the conditional logit. The error components are assumed to be from a normal distribution with zero mean and standard deviation one. When the additional error terms are associated with the alternatives that comprise constraints for wind power generation in contrast to Programme A, the utility functions for the three alternatives are:

$$\begin{aligned} U_{\mathrm{A}} &= \mathrm{ASC_A} + \beta x_{\mathrm{A}} + \varepsilon_{\mathrm{A}} \\ U_{\mathrm{B}} &= \beta x_{\mathrm{B}} + E_{\mathrm{BC}} + \varepsilon_{\mathrm{B}} \\ U_{\mathrm{C}} &= \beta x_{\mathrm{C}} + E_{\mathrm{BC}} + \varepsilon_{\mathrm{C}} \end{aligned} \tag{2}$$

where $\mathrm{ASC_A}$ is the alternative specific constant for Programme A and E_{BC} indicates the error component shared by the two Programmes B and C that constrain wind power development.

Separate models are estimated for both the test and the retest. Additionally, pooled models are obtained by stacking the two databases. A pooled model allows for testing differences in estimated utility parameters across the test and retest using a likelihood-ratio test. Moreover, a pooled model also enables an estimate of the ratio of the scale factor between two data sets (Swait and Louviere 1993). As the scale parameter μ is confounded with the parameter coefficient β it cannot be separately identified in a single data set.[2] Differences in scale between both the test and the retest might be caused by effects such as learning, fatigue, complexity

[2] Whether models such as the G-MNL allow to separately identify the scale parameter is currently debated. Hess and Rose (2012), for example, argue that the attempts presented in the literature to disentangle the two

and consistency as various studies have shown (e.g. Breffle and Rowe 2002; Holmes and Boyle 2005; Campbell et al. 2008; Carlsson et al. 2012). In the present case a learning effect might occur as respondents face the same choice sets a second time. However, since the time span between test and retest is eleven months the learning effect, indicated by a lower error variance in the retest, might be rather weak and not statistically significant. We present two pooled models, one without and one with scale parameter. The scale parameter is explicitly included (in contrast to using a grid-search procedure) and estimated by full information maximum likelihood (Campbell et al. 2008; Olsen 2009). For the subset of choices monitored at the first point in time (the test) the scale parameter is arbitrarily normalised to one while the scale parameter for the subset of choices monitored at the second point in time (the retest) is allowed to vary. All models were estimated in BIOGEME 2.1 (Bierlaire 2003) using Halton draws with 500 replications.

Table 6 reports the estimation results. Overall, both models are statistically significant and the attribute estimates have the expected and same signs in both models. Respondents prefer smaller wind farms compared to Programme A (the future status quo), but both estimates are not statistically significant at the 5 % level in both the test and the retest. Regarding turbine heights respondents are in disfavour of having small (110 m) instead of large turbines (200 m). This estimate is not significant in the test but in the retest indicating a change between both points in time. A change toward medium sized turbines (150 m) is valued positively but this estimate is not statistically significant in any model. Next, the red kite attribute estimates show the same signs and are highly statistically significant in the two models. Reducing the impact of turbines on the red kite is valued positively while increasing the impact on the red kite populations is valued accordingly negatively. Finally, also the distance attribute shows the same signs in all models and is always highly statistically significant. On average, respondents do value higher distances of turbines to residential areas positively. The attribute price is statistically negatively significant in both the test and the retest while the error component is positively significant.

Whether the overall models for the test and the retest differ is investigated through a likelihood ratio test for equality of all model parameters (Swait and Louviere 1993). First, we test for differences among the models for the test, the retest, and the pooled dataset without scale. The chi-square value is 75.76 ($= -2 * [-1444.31–(-711.67 + -694.76)]$, $p < 0.01$). Hence, we reject the hypothesis of parameter equality. We then check for differences when the variance-scale ratio is taken into account. This test comprises the test, retest, and the scale corrected pooled model. In this case the chi-square value is 75.40 ($= -2 * [-1444.13–(-711.67 + -694.76)]$, $p < 0.01$). Again we reject the null hypothesis of equal parameters that means that even when we control for the scale ratio between the two models the parameters are not equal in the test and the retest.

The estimated scale ratio, lambda, reveals the variance of the unobserved factors in the retest relative to that in the test. The estimated ratio is, with a value of 1.08, statistically significantly different from 1. Accordingly, the variance of the unobserved effects in the retest is 14 % lower than in the test. Interpreting the scale parameter as an indicator of the ability to choose (Christie and Gibbons 2011), respondents choices were more consistent during the retest eleven months after they first encountered the choice sets. This means that the data support a weak learning effect that could be attributed to the fact that respondents were more familiar with the measurement instrument.

Footnote 2 continued
components are misguided. According to their findings, the various model specifications, for example presented by Fiebig et al. (2010), are different parameterisations that do allow for more flexible distributions but do not capture scale heterogeneity.

Table 6 Estimates from error component logit models for the test, retest and pooled data

Attribute	Test		Retest		Pooled not scale corrected		Pooled scale corrected	
	Parameter	| t value |	Parameter	| t value |	Parameter	|t value |	Parameter	| t value |
Wind farm: large → medium	0.069	0.68	0.079	0.93	0.069	1.16	0.066	1.16
Wind farm: large → small	0.062	0.69	0.168	1.88	0.120	1.95	0.117	1.98
Turbine heights: 200 m → 110 m	−0.100	1.18	−0.257	2.96	−0.166	2.82	−0.163	2.85
Turbine heights: 200 m → 150 m	0.081	0.98	0.100	1.23	0.082	1.45	0.079	1.45
Red kite: 10 % → 5 %	0.597	6.44	0.498	5.54	0.504	8.18	0.482	6.90
Red kite: 10 % → 15 %	−0.559	5.77	−0.670	6.85	−0.573	8.66	−0.551	7.57
Distance: 750 m → 1,100 m	0.344	4.07	0.264	3.10	0.285	4.93	0.272	4.56
Distance: 750 m → 1,500 m	0.416	4.84	0.222	2.69	0.288	5.01	0.272	4.50
Price	−0.223	5.88	−0.251	6.54	−0.218	8.48	−0.209	7.38
ASC_A	−0.299	0.75	−1.410	2.98	−0.487	1.69	−0.485	1.75
EC_{BC}	3.760	8.59	4.470	7.85	2. 920	10.95	2.810	8.95
μ_{test}						1.000		
μ_{retest}						1.080	7.49[a]	
N observations	860		860		1720		1720	
$Model_K$: Log-L	−944.81		−944.81		−1889.61		−1889.61	
(S)Log-L	−711.67		−694.76		−1444.31		−1444.13	
Pseudo R^2	0.25		0.27		0.24		0.24	

[a] The t value for the scale factor is a t value tested against the null hypothesis H_0: scale = 1

Table 7 Willingness-to-pay estimates (in Euro per month) and confidence intervals based on the models for the test and retest

Attribute	Test	Retest	Poe-test
Wind farm: large → medium	0.31	0.32	0.49
	(−0.47/1.09)	(−0.35/0.98)	
Wind farm: large → small	0.28	0.69	0.22
	(−0.55/1.12)	(−0.06/1.43)	
Turbine heights: 200 m → 110 m	−0.46	−1.05	0.87
	(−1.26/0.33)	(−1.79/−0.31)	
Turbine heights: 200 m → 150 m	0.38	0.41	0.47
	(−0.40/0.33)	(−0.23/1.08)	
Red kite: 10 % → 5 %	2.76	2.03	0.85
	(1.58/3.95)	(1.17/2.89)	
Red kite: 10 % → 15 %	−2.59	−2.74	0.59
	(−3.74/−1.44)	(−3.75/−1.73)	
Distance: 750 m → 1,100 m	1.58	1.07	0.19
	(0.63/2.54)	(0.31/1.82)	
Distance: 750 m → 1,500 m	1.90	0.93	0.04
	(0.98/2.82)	(0.27/1.59)	

Proceeding with the comparison between test and retest, we compare the willingness-to-pay estimates between the test and the retest by employing the complete combinatorial method proposed by Poe et al. (2005). The marginal willingness-to-pay estimates, together with the 95 % confidence intervals based on the Krinsky and Robb (1986) bootstrapping procedure with 1,000 draws, as well as the results of the Poe-Test are reported in Table 7. The complete combinatorial approach indicates only for the distance attribute that describes the move from 750 to 1,500 m a statistically significant difference between test and retest ($\gamma = 4\,\%$). Therefore, the marginal willingness to pay has only changed significantly for the attribute distance when distance is about to be increased from 750 to 1,500 m. In this case the willingness-to-pay estimate from the retest is statistically significantly smaller than the estimate from the test.

In additional models, responses to follow-up questions concerning exposure to the issue of wind power generation since the main survey changes in the personal financial status and perceived effects by the financial crisis were incorporated in the models via interactions with the alternative specific constant as well as with the choice attributes. In all models, the parameter estimates for the interaction terms were not statistically significant. This means that, we do not find any evidence for effects of potential attitude changes, changes in the budget constraint or external events in our study at the level of parametric analysis either.

5 Discussion and Conclusions

This paper presents the first study on test–retest reliability of choice experiments in environmental valuation in which respondents completed five identical choice tasks at two different points in time. The time interval between the test and the retest was eleven months. Based on

a cumulative analysis our findings indicate a fair to moderate test–retest reliability at the level of choice sets, congruent choices per individual, parametric analysis and willingness-to-pay values. Due to a sufficiently long time interval, the presented results can be expected to be robust to a memory or carry-over effect. Moreover, respondents' answers to the follow-up questions indicate no significant effects of attitude changes, changes in respondents' financial status or external events on test–retest results.

The share of congruent choices found in our test–retest study is lower than those shares found in the three studies from health economics (presented in Sect. 2.3). There may be various reasons for the differences between our results and those from the choice experiments conducted in health economics. First, our choice experiment contains two hypothetical alternatives and a no-cost alternative per choice set. This implies a more complex design compared to the health economics studies with two alternatives per choice set. However, the health economics studies included a higher number of attributes and/or a higher number of attribute levels. Further, respondents had to answer a higher number of choice sets in these experiments.

Second, the time span between test and retest in the health economics studies was very different from the time span in the present study. The shortest time span between the test and the retest was almost zero because the first retest was conducted immediately after the test, the longest time span was four months. A retest immediately after the test is a poor measure for test–retest reliability. The rather short time intervals of the health economics studies imply, in general, a higher likelihood of a memory effect compared to our study with an eleven months time span. Third, the choice experiments from health economics solely comprise use values. In contrast, goods valued in environmental studies do not only provide use values but also non-use values. As the good in the present study can also have non-use values, for example, benefits from conserving the red kite, the lower familiarity with goods that provide non-use values might lead to lower test–retest reliability. However, whether the type of benefit, that is, being a use or non-use value or a mixture of both, influences test–retest reliability could only be judged on the basis of a larger number of studies.

As the opportunity to conduct the present test–retest study aroused after the main survey had been carried out, possible requirements of a test–retest study were not considered when the main survey was designed. For example, the present study might be prone to interviewer effects as we were not able to ensure that each respondent was interviewed by the same interviewer in the test and retest. Future studies might try to avoid this source of potential bias although it could be difficult to achieve this taking into account that the same interviewers have to be present at the survey organisation. Another shortcoming is that due to budget constraints only a subgroup of the respondents of the main survey could be interviewed in the retest survey. Thus, only a part of the experimental design applied in the main survey is used in the test–retest study. The part used was, moreover, determined by the respondents who agreed to give another interview and not controlled by the researchers. Thus, future studies should aim at interviewing all respondents of a larger sample twice in order to avoid influences on the test–retest results at this level. Future studies should also seek to present respondents more choice sets as used in this study in order to investigate, for example, learning effects.

Given different levels of comparison between the choices and values at the test and the retest a cumulative test strategy, as employed in this paper, is the most informative. However, it can be argued that a comparison at the level of choices sets is a poor measure of test–retest reliability because it is deterministic. Changes at the level of individual choices might aggregate to non-significant differences on the level of parameter estimates and willingness-to-pay values. It would lead to wrong conclusions if studies solely rely on a test–retest analysis at the level of choice sets. Discrete choice models presume a stochastic process; hence, the analysis

of test–retest reliability based on parametric models and willingness-to-pay estimates is more appropriate. As our results show the interaction of choice experiment designs, individuals' decision process and predictions from probabilistic choice models with respect to measures of test–retest reliability at the corresponding levels (individual, parametric, willingness to pay) is an important task for future research. Future studies should investigate appropriate measures of test–retest reliability in choice experiments and the interrelations of different levels of analysis.

Our results indicate that the design of a choice experiment—the construction of choice sets—affects choice consistency. Future research might take up this finding and investigate more systematically the role of choice sets and choice blocks for test–retest reliability. Further design characteristics that might influence test–retest reliability include the number of alternatives per choice set and the number of choice sets presented to a respondent. Furthermore, additional information, for example on changes of attitudes between test and retest, should be recorded to gain more insights into how potential non-congruence of choice behaviour over time can be explained. In the present study such variables did not affect test–retest results, supporting the interpretation of fair to moderate test–retest reliability that is not affected by attitude change or external events. Future test–retest studies that rely on identical choice sets per individual and comprise longer time intervals might also indicate whether and to what extent preference changes occur over time. Particularly a parametric analysis taking into account differences in scale between test and retest can help to differentiate effects of preference changes from other effects such as learning, fatigue and complexity. In our study the results suggest a small learning effect.

The results of the test–retest studies using choice experiments from health economics and the results of our study reveal a range of overall fair to substantial test–retest reliability. Moreover, given a comparable range of time intervals, the determined reliability of the present choice experiments is well in the range of test–retest results reported in contingent valuation studies. Thus, no striking differences between choice experiments and contingent valuation can be observed to date. This is especially noteworthy since choice experiments are generally expected to be much more cognitively demanding than contingent valuation studies. Nevertheless, it has to be stressed that these conclusions regarding test–retest reliability are derived from a very low number of studies, three from health economics and one from environmental economics. Hence, the present study is only a first step and future studies should further investigate the test–retest reliability of choice experiments. As choice experiments are used more and more often in benefit transfer it is important to know more about their reliability. So far most environmental valuation studies simply assume preference stability over time when results are used later on to inform decision makers.

Acknowledgment We are especially grateful to a reviewer who drew our attention to crucial issues regarding the definition and measurement of test–retest reliability of choice experiments. Also, we would like to acknowledge the comments made by Riccardo Scarpa (Associate Editor) and Wojtek Przepiorka. Finally, we would like to thank Christian Vossler for valuable suggestions made as a discussant of a previous version of this paper at the 4th World Congress of Environmental and Resource Economics 2010 in Montreal, Canada. Funding for this research, which was part of the project 'Strategies for sustainable land use in the context of wind power generation' (Fkz. 01UN0601A, B), was provided by the Federal Ministry of Education and Research in Germany.

Appendix

Table 8 Example of a choice set

Wind power in Nordhessen until 2020			
	Programme A	Programme B	Programme C
Size of wind farms	Large farms	Small farms	Large farms
Maximum height of turbines	200 m	110 m	110 m
Impact on red kite population	10 %	5 %	10 %
Minimum distance from residential areas	750 m	1,100 m	1,500 m
Monthly surcharge to power bill beginning in 2009	€ 0	€ 6	€ 1
I choose	☐	☐	☐

References

Ary D, Jacobs LC, Sorensen C, Razavieh A (2009) Introduction to research in education, 8th edn. Wadsworth, Belmont

Bierlaire M (2003) BIOGEME: a free package for the estimation of discrete choice models, presented at the 3rd Swiss transportation research conference, Ascona

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

Bowker AH (1948) A test of symmetry in contingency tables. J Am Stat Assoc 43(244):572–574

Boxall P, Adamowicz WL, Moon A (2009) Complexity in choice experiments: choice of the status quo alternative and implications for welfare measurement. Aust J Agric Resour Econ 53(4):503–519

Breffle WS, Rowe RD (2002) Comparing choice question formats for evaluating natural resource tradeoffs. Land Econ 78:298–314

Brouwer R, Bateman IJ (2005) Temporal stability and transferability of willingness to pay for flood control, and wetland conservation. Water Resour Res 41(3):1–6

Bryan S, Gold L, Sheldon R, Buxton M (2000) Preference measurement using conjoint methods: an empirical investigation of reliability. Health Econ 9(5):385–395

Campbell D, Hutchinson WG, Scarpa R (2008) Incorporating discontinuous preferences into the analysis of discrete choice experiments. Environ Resour Econ 41(3):401–417

Carlsson F, Mørkbak MR, Olsen SB (2012) The first time is the hardest: a test of ordering effects in choice experiments. J Choice Model (forthcoming), Gothenburg

Christie M, Gibbons J (2011) The effect of individual 'ability to choose' (scale heterogeneity) on the valuation of environmental goods. Ecol Econ 70:2250–2257

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

Ferrini S, Scarpa R (2007) Designs with a priori information for nonmarket valuation with choice experiments: a monte carlo study. J Environ Econ Manag 53(3):342–363

Fiebig D, Keane M, Louviere J, Wasi N (2010) The generalized multinomial logit: accounting for scale and coefficient heterogeneity. Mark Sci 29(3):393–421

Guttman L (1945) A basis for analyzing test–retest reliability. Psychometrika 10(4):255–282

Hensher DA, Rose JM, Greene WH (2005) The implications of willingness to pay of respondents ignoring specific attributes. Transportation 32(3):203–222

Hensher DA, Jones S, Greene WH (2007) An error component logit analysis of corporate bankruptcy and insolvency risk in Autralia. Econ Rec 83(260):86–103

Hess S, Rose JM (2012) Can scale coefficient heterogeneity be separated in random coefficient models? Transportation (online 1. April 2012)

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

Johnson FR, Kanninen B, Bingham M, Özdemir S (2007) Experimental design for stated choice. In: Kanninen B (ed) Valuing environmental amenities using stated choice studies. Springer, Dordrecht, pp 159–202

Jorgensen BS, Syme GJ, Smith KM, Bishop BJ (2004) Random error in willingness to pay measurement: a multiple indicators, latent variable approach to the reliability of contingent values. J Econ Psychol 25(1):41–59

Kaplan RM, Saccuzzo DP (2008) Psychological testing: principles, applications, and issues, 7th edn. Wadsworth, Belmont

Krinsky I, Robb AL (1986) On Approximating the statistical properties of elasticities. Rev Econ Stat 68(4):715–719

Landis JR, Koch GG (1977) The measurement of observer agreement for categorial data. Biometrics 33(1):159–174

Lusk JL, Norwood FB (2005) Effect of experimental design on choice-based conjoint valuation estimates. Am J Agric Econ 87(3):771–785

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

Meyerhoff J, Ohl C, Hartje V (2010) Landscape externalities from onshore wind power. Energy Policy 38(1):82–92

Olsen SB (2009) Choosing between internet and mail survey modes for choice experiment surveys considering non-market goods. Environ Resour Econ 44(4):591–610

Olsen SB, Lundhede T, Jacobsen J, Thorsen B (2011) Tough and easy choices: testing the influence of utility difference on stated certainty-in-choice in choice experiments. Environ Resour Econ 49(4):491–510

Poe GL, Giraud KL, Loomis JB (2005) Computational methods for measuring the difference of empirical distributions. Am J Agric Econ 87:353–365

Ryan M, Netten A, Skatun D, Smith P (2006) Using discrete choice experiments to estimate a preference-based measure of outcome—an application to social care for older people. J Health Econ 25(5):927–944

Scarpa R, Ferrini S, Willis K (2005) Performance of error component models for status-quo effects in choice experiments. In: Scarpa R, Alberini A (eds) Applications of simulation methods in environmental and resource economics. Springer, Dordrecht, pp 247–273

Scarpa R, Campbell D, Hutchinson WG (2007a) Benefit estimates for landscape improvements: sequential Bayesian design and respondents' rationality in a choice experiment. Land Econ 83(4):617–634

Scarpa R, Willis K, Acutt M (2007b) Valuing externalities from water supply: status quo, choice complexity and individual random effects in panel kernel logit analysis of choice experiments. J Environ Plan Manag 50(4):449–466

Skjoldborg US, Lauridsen J, Junker P (2009) Reliability of the discrete choice experiment at the input and output level in patients with rheumatoid arthritis. Value Health 12(1):153–158

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

Yu CH (2005) Test–retest reliability. In: Kempf-Leonard K (ed) Encyclopedia of social measurement, vol 3 P–A. Elsevier, Amsterdam, pp 777–784