The Australian Journal of Agricultural and Resource Economics

Journal of the Australian Agricultural and Resource Economics Society

Australian Journal of Agricultural and Resource Economics, 59, pp. 375–392

A within-sample investigation of test–retest reliability in choice experiment surveys with real economic incentives

Morten Raun Mørkbak and Søren Bøye Olsen[†]

In this paper, we investigate the level of agreement between respondents' choices in identical choice sets in a test–retest choice experiment for a market good with real economic incentives, thus investigating whether the incentivised CE method can be reliable and stable over time. Besides comparing choices, we also test for differences in preferences and error variance when a sample of respondents is given the exact same questionnaire twice, with a time lag of 2 weeks in between. Finally, we examine potential reasons and covariates explaining the level of agreement in choices across the 2 weeks. Across four different tests, we find very good agreement between the two choice experiments – both with respect to overall choices and with respect to preferences. Furthermore, error variances do not differ significantly between the two surveys. The results also show that the larger the utility difference in a choice task, the larger the probability that the respondent will choose the same alternative in the retest. Moreover, the results show that the longer time respondents take to answer the 12 choice sets in the retest, the lower the probability that the respondent will choose the same alternatives in the retest as they did in the test.

Key words: choice experiments, market good, real incentives, reliability, test–retest.

1 Introduction

Stated preference (SP) methods have been widely used for assessing potential consumer demand prior to actual market introduction of new products. By asking survey participants to act on a fictitious market for the good in question, preferences and thus potential demand for the good can be elicited by observing their behaviour on the hypothetical market. However, the credibility and validity of SP surveys has been questioned for many years. As a result, a substantial body of methodological research has accumulated over the last 20 years seeking to thoroughly explore and investigate the credibility and validity of the stated preference methods.

The focal issue of concern in this paper is that of validity. One measure of the credibility is the temporal reliability, that is, 'Does an individual respondent provide identical answers to identical questions over time, assuming that preferences are stable?' If responses over time are not

[†] Morten Raun Mørkbak (e-mail: mrm@sam.sdu.dk) is at Department of Business and Economics, COHERE, University of Southern Denmark, Odense, Denmark. Søren Bøye Olsen is at Institute of Food and Resource Economics, University of Copenhagen, Copenhagen, Denmark.

doi: 10.1111/1467-8489.12067

reasonably stable, the reliability of stated preference surveys is disputable. In relation to the example above, if the producer, based on a SP survey-based demand revelation, decides to go ahead and introduce the new agricultural product in the market, it would obviously be crucial that the estimated demand is a reliable predictor of actual demand when the good is actually introduced in the market some time later, depending on the time required to produce the good. Moreover, if SP surveys are not reliable, the use of benefit transfer which by definition involves transferring values from one point in time to another would seem particularly questionable. In the stated preference literature, this issue of temporal inconsistency has been investigated within several contexts – both within different methodologically contexts and within different policy contexts (e.g. McConnell *et al.* 1998; Bryan *et al.* 2000; Miguel *et al.* 2002; Brouwer and Bateman 2005; Shiell and Hawe 2006; Ryan *et al.* 2006; Skjoldborg *et al.* 2009; Liebe *et al.* 2012; Boman *et al.* 2011; Rigby and Burton 2011).

One aspect of the concerns about credibility and validity of stated preference methods relates to the hypothetical nature of the methods. The hypothetical nature of stated preference methods such as the contingent valuation method (CV) and choice experiments (CE) often gives rise to different types of biases which are typically collected under one umbrella term, namely hypothetical bias (see, e.g., List *et al.* 2006 and Lusk and Schroeder 2004). The efforts to reduce/eliminate hypothetical bias have taken many different directions, such as the introduction of budget reminders (Mitchell and Carson 1989) or cheap talk scripts (e.g. Cummings and Taylor 1999; Carlsson *et al.* 2005). Hypothetical bias has also been sought ameliorated by introducing real economic incentives in the otherwise hypothetical settings of stated preference methods (see, e.g., Lusk and Schroeder 2004; Alfnes *et al.* 2006; Scarpa *et al.* 2013). As this approach seems particularly appealing from a theoretical point of view in terms of increasing incentive compatibility, more and more stated preference surveys incorporate real economic incentives when possible.

Despite this increased interest in and acceptance of valuation methods with real economic incentives, and despite the attention paid to temporal reliability in the SP literature mentioned above, to our knowledge, no one has yet specifically investigated the temporal reliability of CE surveys in settings with real economic incentives. With this paper, we aim to contribute to this gap in the literature by investigating the level of agreement between respondent choices in identical choice sets in a repeated choice experiment, that is, a test–retest experimental setting. In the present survey, the chosen measure of reliability is when respondents choose the same alternative in the retest as they did in the test survey. It may be argued that this choice of reliability definition could be less restrictive. For instance, if two alternatives are close in utility, that is, utility balance is high, considering the underlying random utility framework it could be considered less unreliable if the respondent does not choose identical alternatives in the test and retest, than if the two

alternatives were far apart in utility, that is, utility balance low.[1] Nevertheless, given the purpose of this paper, the strict definition of reliability is used here, implying that the two examples above will be treated identically, namely as being in violation of our reliability definition.

Besides comparing choices, we also test for an effect on preferences and error variance when a sample of respondents is given the exact same questionnaire twice, with a time lag of 2 weeks in between. Finally, we examine potential reasons and covariates explaining the level of agreement in choices across the 2 weeks.

2 Previous test–retest studies

Preference stability in stated preference surveys with and without repeated questions has received considerable attention in the literature. However, perhaps as could be expected, no consistent pattern has emerged regarding the stability of preferences over time. Previous test–retest studies have examined the reliability of the estimates within both the CVM and the CE framework, respectively. Since the present paper investigates this within the CE framework, the following review of the literature only reports on findings from similar frameworks. Furthermore, in the literature reported below, reliability is defined as in the present paper, that is, intrapersonal identical choices in the test and the retest are considered reliable.

Bryan *et al.* (2000) completed a test–retest survey concerning health care using CE. The retest was conducted with a time lag of approximately 2 weeks. At the input level or the choice task level, the level of agreement was found to be good according to the kappa statistics of 0.65, and when testing on the output level – the preference level – the results were also overlapping.

In the light of Bryan *et al.* (2000), Miguel *et al.* (2002) investigated the effect of experience on stability of preferences within health care in a CE setting. The results showed that preferences did not change due to experience – the no-experience respondents actually did change their preferences in the retest survey, even though the nonparametric tests between the test and the retest showed a good level of agreement between both groups.

Ryan *et al.* (2006) tested the test–retest reliability in a healthcare context applying the CE setting. The retest was carried out on 47 respondents who completed a second questionnaire between 11 and 60 days after the first one. The results showed a kappa coefficient of 0.64, suggesting a good strength of agreement. Moreover, the LR test of equality of preferences could not be rejected, implying that the parameter estimates were reliable between the two periods.

Conjoint reliability over time was tested by Skjoldborg *et al.* (2009) focussing on temporal reliability, where the same instrument and the same respondents were used three times over a 4-month period. At the choice task

[1] We would like to thank a reviewer for bringing this to our attention.

level, the level of agreement between the three surveys was fair, with agreement levels between 75 per cent and 87 per cent, and with respect to WTP estimates, the results were the same – no differences were found between the three surveys.

Also within the CE framework, Liebe *et al.* (2012) conducted a test–retest study on the valuation of landscape externalities from onshore wind power. With a time interval of 11 months between the test and the retest, and with the sample constraint on the retest sample that they should have participated in the test sample, they found a fair test–retest reliability on approximately 60 per cent of individual choices, but when it came to the WTP estimates there were found several significant differences (4 out 9 parameters differed).

Liekens *et al.* (2012) report on a test–retest CE involving a time lag of 12 months. They examine temporal reliability within simple choice occasions, as well as comparing preference parameters, WTP and error variance. The results show that choices change in almost 50 per cent of the choice situations. With regard to the preference parameters, the results show that neither preference parameters nor marginal willingness to pay estimates differs significantly across time. Finally, they find that error variance does change in terms of choices being less consistent in the retest, 1 year later.

Finally, Rigby and Burton (2011) conducted a study on intertemporal choice consistency in CE, with a time lag of 6 months using a repeated subsample of respondents. They found a choice consistency within 64 per cent of the choice occasions, but no differences in the parameter estimates over time. Moreover, they examined the respondent-specific characteristics explaining the consistency. These results showed that respondents' capability and commitment to the survey process and the complexity of the choice task were of significance when explaining the consistency.

3. Method and data

The empirical survey considers consumers' preferences for an everyday, low-budget-share market good, namely apples. The CE method was employed in an experimental set-up where respondents were provided with a real economic incentive in order to increase truth-telling and have respondents make choices that as closely as possible reflected their 'true' preferences. The procedure was the following: the respondents were given a fixed show up fee just for coming to the survey room. The rest of the payment was determined depending on how they chose during the choice experiment. The following description was provided to the respondents following Chang *et al.* (2009):

'After everyone completes all 12 shopping scenarios, we will ask for a volunteer to draw a number (1 to 12) from a hat to determine which shopping scenario will be binding. In the hat are numbers 1 through 12. If the number 1 is drawn then the first shopping scenario will be binding. If the number 2 is drawn the second shopping scenario will be binding, and so on. For the binding scenario, we will look at the product you have chosen, give you your

chosen product, and you will pay the listed price in that scenario. You will be given a value ticket of 45 Danish Kroner (DKK), which you should use for the purchase. The most expensive alternatives cost DKK 45. If you choose a cheaper alternative you will be given the remaining money. Although only one of the 12 shopping scenarios will be binding there is an equal chance of any shopping scenario being selected as binding, so think about each answer carefully.'

The procedure of randomly drawing one binding choice set for all individuals follows the procedure of Lusk and Schroeder (2004) and Alfnes *et al*. (2006). As all choice sets have an equal chance of being realised as a result of the random draw, this procedure should theoretically ensure incentive compatibility across the entire sequence of choice sets.

The attributes and levels used in the CE design (displayed in Table 1) were identified through focus group testing and a pilot test. The experimental design used is a D-efficient fractional factorial design with fixed priors resulting in 12 different choice sets in total. The software Ngene was used to generate the alternatives and the choice sets (Rose *et al*. 2009). The 12 choice sets each consist of two generic alternatives plus a status quo alternative defined as a standard bag of low-priced apples (conventional mixed colour of sour and mealy apples produced outside EU at a price of DKK 7). The first data collection took place in March 2011 where 25 face-to-face group interviews were conducted. The exact same 25 respondents were invited to participate in an identical survey 2 weeks later. The time span of 2 weeks was chosen so that the risk of external events causing respondents to change attitudes and preferences between the test and retest was minimised. However, using such a relatively short time span raises the issue of possible carry-over effects from the test to the retest, that is, respondents simply remembering their choices from the first survey and then answering the same in the retest in order to appear consistent. To minimise such carry-over effects, respondents were not explicitly told that the second survey was identical. They were simply informed that the second survey considered the same good and that it would have some major resemblances with the first survey, but that they were to answer according to their preferences now. Furthermore, the full questionnaire had 127 questions. For a respondent to

Table 1 The attributes and levels used in the CEs.

Characteristics	Levels
Type of production	Conventional, organic
Origin	Locally produce (Danish), Danish produce, European produce (not Danish), produced outside Europe
Colour of apples	Red, green, yellow, mix of colours
Taste of apples	Sweet and crunchy, sweet and mealy, sour and crunchy, sour and mealy
Price (DKK per kg)	7, 15, 25, 45

Note: DKK 10 ~ EUR 1.34.

remember their answers to all these questions, including the 12 choice task questions placed towards the end of the questionnaire, for 14 days would seem to represent a significant cognitive challenge. Based on informal feedback from respondents after the last group interview, it seemed that few respondents had realised that it was exactly the same questionnaire (rather than just a very similar one), and few were actually able to remember their answers from the first survey. To further control for this, we incorporated an extra follow-up question after the CE regarding this specific issue. Apart from this, the two questionnaires were kept identical in order to control for framing effects. Respondents had to answer the questionnaire individually, and only the randomly binding choice set was announced in the group after all had finished the questionnaire. All respondents were recruited from a consumer panel of people living in Copenhagen. The sample consists of a majority of women, with an average of age of 50 and with less than one child still living at home (see appendix Table A1). In relation to the efficiency of the used design, the posterior D-error measure has been computed according to the approach suggested by Scarpa and Rose (2008). The posterior D-error measures for the two CEs were 0.0213 and 0.0199, respectively, suggesting no difference between the test and retest – and both major improvements relative to the prior D-error of 0.612, though computed for a MNL model. Finally, the design had an S-error of 25, indicating that a sample of 25 respondents should be sufficient to obtain significant model parameters (Rose *et al*. 2009).

4. Testing stability

Stability of choices and preferences is examined both nonparametrically and parametrically. The nonparametric approaches assess the level of agreement between the choices at the two different points in time, both with and without controlling for the issue of chance. The parametric approaches assess both comparisons of the overall preference structure, the individual WTP estimates and an analysis of potential factors explaining the level of agreement.

4.1. Test 1: Gross level of agreement

The gross level of agreement is a simple and direct analysis of testing the stability of choices over time. The analysis shows the proportion of respondents reversing their choice between the two questionnaires.

4.2. Test 2: Gross level of agreement corrected for chance

One limitation of test 1 is that it does not take into account the fact that the respondents might choose the same alternative in the two CE exercises by chance, which would bias the level of agreement. By estimating Cohen's κ coefficient, we allow for the effect of chance (Cohen 1968) and make the same comparison as done in test 1.

4.3. Test 3: Parametric test of equality of preferences – the LR test

The difference in preferences between the two time periods is initially examined through a likelihood ratio (LR) test for equality of all model parameters, including the scale parameters (Swait and Louviere 1993). If the test reveals that there are no significant differences in the overall preference structure it will imply that the preferences are stable. With respect to the scale parameter, the *a priori* expectation is that either there are no differences in scale between the two time periods or alternatively the scale is larger in the second period.

4.4. Test 4: Test of equality of WTP estimates

In addition to comparing the estimated parameters and relative scale parameters between models, we also compute and compare unconditional marginal WTPs for each time split. The advantage of such a comparison is that the scale parameters cancel from the expression and we can thus directly compare mean WTP estimates between models in the two periods, thus providing the reader with a direct measure in monetary terms of any potential differences between the two periods.

4.5. Test 5: Hit rate model

Given our within-subject design, we can make a number of comparisons. Following Rigby and Burton (2011) and Carlsson *et al*. (2012), we use a binary probit model in order to explain 'hit success', where the dependent dummy variable takes a value of one if the respondents makes the same choice in the identical choice sets in the first and the last sequence and zero otherwise. As explanatory variables in the random effects binary probit panel model, we use (i) socio-demographic characteristics (gender, age and household income); (ii) respondents self-reported certainty statements[2] ; (iii) utility difference at choice set level as a proxy for the complexity of a choice set; (iv) an ordering/learning effect which is captured by a dummy variable for the first choice set with the remaining 11 choice sets as the baseline and finally; and (v) response time.

5. Econometrics

The underlying theory of CE is based on Lancaster's consumer theory (Lancaster 1966) and random utility theory (Luce 1959; McFadden 1974). In the former, consumer preferences were defined in relation to bundles of characteristics and the demand for goods was a derived demand. Consumption was the activity of extracting characteristics from goods (Gravelle and

[2] Respondents were asked to state how certain they are after the choice experiment with respect to institutional and value uncertainty, respectively.

Rees 1992). In the analysis underlying tests 3 and 4, we apply a standard random utility model (McFadden 1974), where the utility of alternative j for individual i in choice set k is specified as

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

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

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

where μ is a scale parameter that is inversely proportional to the error variance. The coefficients (β) in the econometric models are usually expressed in their scaled form ($\beta = \mu\beta^*$), where the scale parameter μ and the 'original' coefficients β^* are confounded. Hence, the estimated parameter β indicates the effect of each observed variable relative to the variance of the unobserved factors (Train 2003).

In the present case, an error component logit model representation of mixed logit was found suitable.[3] Since the status quo alternative was constant and presented as a standard bag of cheap apples with (what we *a priori* expected to be) the least preferred quality attribute levels, we deemed it important to account for potential status quo effects in our econometric model. Following Scarpa *et al*. (2005), an alternative specific constant (ASC) is specified for the status quo alternative in order to capture the systematic component of a potential status quo effect. Furthermore, an error component additional to the usual Gumbel-distributed error term is incorporated in the model to capture any remaining status quo effects in the stochastic part of utility. The error component, which is implemented as a zero-mean normally distributed random parameter, is assigned exclusively to the two non-status quo alternatives. Thus, correlation patterns in utility over these alternatives are induced (Brownstone and Train 1999; Herriges' and Phaneuf 2002; Scarpa *et al*. 2005, 2008).

As the utility function is assumed to be linear in the attributes, the marginal WTPs for the attributes which are investigated in test 4 are the ratio between the parameter of the attribute and the cost parameter in the utility function (2), such that:[4]

[3] Employing random parameter error component logit (ECL) models might be more informative if one aims for knowledge about heterogeneity in preferences (Greene and Hensher 2007; Scarpa *et al*. 2007, 2008). However, when testing these models on our data, we found only very limited preference heterogeneity and no overall improvement of model fit when adjusting for added parameters.

[4] The standard errors of the WTP are estimated using the Delta method (Greene 2003).

$$\text{WTP} = -\frac{\text{Attribute parameter}}{\text{Cost parameter}} \tag{3}$$

We have used the software package Biogeme (Bierlaire 2003) to estimate the econometric models. In all models, a panel specification capturing the repeated choice nature of the data in terms of the 12 choice sets per respondent is used. The models are estimated with simulated maximum likelihood using Halton draws with 300 replications.

For test 5, in line with Olsen *et al.* (2011), we use the estimated model in equation (2) to calculate the expected aggregate utility of each alternative for each individual and then calculate the expected utility difference, UD, between the alternative chosen, k, and the best alternative to that (either l or m), that is, for each choice set:

$$\begin{aligned} \text{UD} &= E(u_{ki}(x_{ki}, \varepsilon_{ki})) - \max\{E(u_{li}(x_{li}, \varepsilon_{li})); E(u_{mi}(x_{mi}, \varepsilon_{mi}))\} \\ &= \hat{\beta}'_i x_{ki} - \max\left\{\hat{\beta}'_i x_{li}; \hat{\beta}'_i x_{mi}\right\} \end{aligned} \tag{4}$$

That is, the utility of each alternative is calculated by multiplying the estimated utility weights with the corresponding attribute levels. This utility difference is used as an explanatory variable in a standard random effects binary probit panel model to identify explanatory variables for hit success. The binary probit model is an index model specified as the conditional probability that a binary response variable will take the value one:

$$P(y = 1|x) = G(x'\beta) \equiv p(x) \quad \in [0, 1] \tag{5}$$

Here, y is the binary response variable, x is a vector of explanatory variables, β is a parameter vector, and G is a function mapping the linear index $x'\beta$ into the response probability. In our analyses, the response variable, y, takes the value one if the respondent has made the same choice in the two surveys and zero otherwise. The utility difference variable enters the model as an explanatory variable in x. Remaining explanatory variables in x have been chosen on account of *a priori* expectations of their potential impact on hit success. The probit model assumes that the unobservable error terms are normally distributed with mean zero and that the unobserved heterogeneity for the random variables also follows a normal distribution.

6. Results

As a first indication of stability of answers, respondents were asked whether they believed that they had changed their answers from the test to the retest. Only 8 per cent of the respondents stated that they had chosen different alternatives in the choice task in the retest than in the test, while 24 per cent answered that they had not changed their answers. The remaining 68 per cent

of the respondents did not know. Examining the numbers in more detail for the respondents who stated that their choices were stable, their actual choices confirmed this; there was a 90 per cent correspondence between choices made in the test and the retest. However, there was an 83 per cent correspondence between choices made in the test and the retest by the respondents who stated that they had in fact changed their answers. Though not strong evidence, this would suggest that respondents have found it difficult to remember their choices from the test to the retest survey. Thus, any 'carry-over' effect from the test to the retest is likely to be relatively limited.[5]

The gross level of agreement test (test 1) showed that individual respondents chose identical alternatives in the two interviews in 78.7 per cent of the choice tasks. Of these, 28 per cent of the respondents always chose the same alternative in the retest as in the initial test, whereas 52 per cent of the respondents chose the same alternative in at least 11 out of the 12 choice tasks and 56 per cent of the respondents at least 10 out of the 12 choice tasks, respectively. As mentioned above, the next test takes the effect of chance into account (test 2). The probability of a random agreement is 0.45, which further provides us with a Cohen's κ coefficient[6] of 0.62, suggesting a 'good agreement' between the two surveys. Hence, tests 1 and 2 do not suggest any instability of choices from the test to the retest.

6.1 Parametric analyses

Moving on to the parametric analysis enabling tests 3 and 4, Table 2 and Table A2 (in the appendix) shows the LR statistics for the four error component models[7] – one for each of splits 1(test) and 2(retest), one for the pooled data set not accounting for potential differences in scale and finally one pooled model where potential differences in scale are accommodated (the entire models are shown in appendix Table A2). As can be seen from the table, the adjusted pseudo R^2 values ranging between 0.43 and 0.49 suggest a very good fit of the models to the data.

This test 3 involves models (i)–(iii). Comparing the pooled model in (iii) with the two separate models (i) and (ii), the value of the chi-squared test statistic is found to be 16.50 – which means that we cannot reject the hypothesis of equal parameters at the standard 5 per cent level of statistical significance (critical value at 5 per cent and 10 df. is 18.31). Since model (iii) not only constrains preference parameters to be equal but also scale factors,

[5] This does not say that respondents did not try to remember their earlier choices – just that they did not succeed in doing so.

[6] Cohen's κ coefficient is defined as $\kappa = \frac{p_0 - p_t}{1 - p_t}$, where p_0 is the observed agreement, and p_t is the agreement that you would get just by chance.

[7] As can be seen from comparing the attributes in Table 1 and the estimates in Table A2, for reasons of simplicity, we have merged some of the attribute levels since they were not significantly different from each other. Moreover, we have tested for taste heterogeneity, but found none. While this would suggest homogeneous preferences, it should be noted that the sample size is relatively small.

Table 2 LR statistics for the EC models

	Model (i) Split 1 – test	Model (ii) Split 2 – retest	Model (iii) Pooled without scale correction	Model (iv) Pooled with scale correction
LL	−178.165	−159.754	−346.171	−345.257
Pseudo R^2	0.429	0.485	0.46	0.46
LR test statistics			16.504	14.676

this joint restriction is accepted under the test, and it is essentially redundant to test any further for differences in scale or error variance. Nevertheless, for completeness, we proceed with a pooled model where we account for a potential difference in scale parameters across the test and the retest, but where the preference parameters are restricted to be the same across the two sequences (model iv). Recall that the scale parameter is inversely proportional to the standard deviation of the error term in our specification (Swait and Louviere 1993). Our results confirm that the error variance between the test and the retest does not differ significantly. Hence, test 3 suggests that preferences as well as error variance are stable when moving from the test to the retest.

While these results are in line with some previous studies looking at the stability of preferences in CE (e.g. Carlsson and Martinsson 2001; Hanley *et al*. 2002), they are in contrast to others (e.g. Holmes and Boyle 2005; Day *et al*. 2012). However, the majority of these previous studies testing for differences in WTP are between-sample tests. In this paper, we have the opportunity to provide this particular type of within-sample test for differences in WTP in identical choice sets. In order to further test for this (test 4), we report the unconditional marginal mean WTP for each attribute and the ASC obtained on the basis of the indirect utility parameter estimates in Table 2 for models (i) and (ii). Table 3 presents the results. It is evident that the mean WTP estimates are far from significantly different across the

Table 3 Unconditional WTP estimates based on models (i) and (ii) (test 4)

	Split 1 – test		Split 2 – retest		*t*-value
	WTP	Std. Err. (WTP)	WTP	Std. Err. (WTP)	
ASC SQ	−149.30	85.35	−93.95	65.15	-0.52^{ns}
Organic produce	39.81	19.28	38.67	22.54	0.04^{ns}
Local produce within Denmark	58.60	25.77	58.87	26.91	-0.01^{ns}
Danish produce	−32.09	23.48	−39.52	31.21	0.19^{ns}
Green coloured apples	18.37	11.39	36.05	25.66	-0.63^{ns}
Yellow coloured apples	72.56	32.97	97.98	53.02	-0.41^{ns}
Red coloured apples	40.88	23.16	73.79	42.54	-0.68^{ns}
Sweet and mealy apples	22.65	19.27	44.35	35.62	-0.54^{ns}

Notes: ns indicates insignificance.

two surveys, and thus, test 4 also suggests that preferences are stable from the test to the retest.

Another interesting question is whether the change in preferences can be traced to certain choice sets or a certain part of the order of the choice sets. Given our within-subject design, we can make a number of comparisons. Test 5 models the 'hit success', where the dependent dummy variable takes a value of one if the respondent makes the same choice in the identical choice sets in the test and the retest, and zero otherwise. Only parameters with significant distributions are included. The results are presented in Table 4.

The only significant socio-demographic effect is that respondents with larger income tend to have a smaller/lower 'hit' rate than their lower-income counterparts. This is somewhat surprising since we had expected the opposite. One could argue that since high income groups are less sensitive to marginal

Table 4 Hit rate random effect binary probit panel model (test 5)

	Description	Coefficient	Std. Err.	*t*-value
Female	Dummy = 1 if respondent is a female, else 0 (mean 0.79)	0.149	0.388	0.380^{ns}
Age	Age in years (mean 50.5)	0.001	0.014	0.100^{ns}
Household income	1 if below DKK 100,000; and10 if above DKK 900,000 (mean 5.17)	−0.059	0.034	−1.720*
Institutional uncertainty	1 if totally incomprehensible; and5 if easily comprehensible, stated in initial survey (mean 3.54)	0.021	0.326	0.060^{ns}
Value uncertainty	1 if very uncertain; and 5 if very certain, stated in initial survey (mean 3.29)	0.071	0.280	0.250^{ns}
Utility Diff.	Estimated average utility difference (mean 0.29)	2.280	0.384	5.930***
Time	Response time for answering the 12 choice sets in the retest (mean 196.83)	−0.006	0.004	−1.800*
Status Quo	1 if respondent chose the status quo in the initial survey (mean 0.06)	1.520	1.116	1.360^{ns}
CS 1	Dummy = 1 if choice set 1, else 0 (mean 0.056)	0.351	0.533	0.660^{ns}
Constant		1.665	1.481	1.120^{ns}
Parameters for dists. of random parameters				
Utility Diff.	1.726	0.334	5.160***	
Institutional uncertainty	0.127	0.042	3.040***	
Household income	0.055	0.018	3.020***	
Observations	300			
Respondents	25			
LL	−89.33			
Pseudo R^2	0.12			

Notes: ***, ** and * indicate the 1%, 5% and 10% levels of significance – ns indicates insignificance.

changes in income, the price attribute tends to be of less importance than the other attributes compared to the case for the low income group. This makes the choice of alternative more random for the high income group, since the price attribute is less salient for the high income group. Maybe also slightly surprising, we find no evidence of learning effects affecting the hit rate. We had expected that learning effects especially associated with the first choice set in the first round would lead to a reduced chance of choosing the same alternative in choice set 1 in the retest. Another *a priori* expectation was that respondents who felt relatively more certain about their choices than others would also tend to exhibit more stable choices across the test and the retest survey. However, the results show that there is no significant mean effect of respondents stating that they feel relatively certain both in terms of value/preference certainty and institutional certainty, but as the random component shows, there appears to exist some heterogeneity with respect to the institutional certainty.

The significant impact of utility difference reveals that the larger the utility difference in a choice task, the larger the probability that the respondent will choose the same alternative in the retest. This conforms to our expectations since making a choice in a choice set with large utility difference will, *ceteris paribus*, be easier than choosing from a choice set where utility is more balanced. As described above, three parameters were estimated as random parameters. The significant estimate for the random effect associated with the utility difference measure suggests that the impact of utility difference is heterogeneous across respondents. This implies that for some respondents, the utility difference has no impact on the probability of making the same choice in the retest as they did in the test. Such individuals could potentially be individuals using decision heuristics or having lexicographical preferences – though not trading off the attribute (levels) between the alternatives.

Moreover, the results show that the longer time a respondent takes to answer the 12 choice sets in the retest, the lower the probability that the respondent will choose the same alternative as in the test. This is contrary to our *ex ante* expectations, since one could expect that more certain respondents (respondents using more time) would have a larger probability of choosing the same alternative in the retest as they did in the test. One potential explanation could be that respondents use more time answering simply because they are – unsuccessfully – trying to remember what alternative they chose last time. Unfortunately, our experimental set-up does not allow us to test this carry-over explanation, and it thus remains speculative.

7. Discussion and conclusion

The credibility of stated preference surveys has been seriously questioned in the literature for many years. As a response to this criticism, research into the credibility of stated preference methods has emerged. One measure of the

credibility is the reliability, that is, 'Do respondents answer the same to identical questions at different points in time, assuming that preferences are stable?' This is what is examined in the present paper. More specifically, we investigate the level of agreement between the two time periods, as well as the effect on preferences and error variance when the same sample of respondents is introduced to an identical CE questionnaire with a time lag of 2 weeks between surveys and using an incentivised setting in both cases. Finally, we examine potential reasons and covariates explaining the level of agreement.

Across four different tests, we find very good agreement between the two choice experiments – both with respect to overall choices and with respect to preferences – and we conclude that the CE method provides fairly reliable results in the given empirical case. Looking into what might explain the level of agreement between choices in the test and retest CE surveys, we find that the larger the utility difference in a choice task, the larger the probability that the respondent will choose the same alternative in the retest. Moreover, the results show that the longer time respondents take answering the 12 choice sets in the retest, the lower the probability that the respondent will choose the same alternative in the retest as they did in the test.

Taking our results at face value, we can conclude that CE results obtained for a marketed good in a real incentive setting can indeed be reasonably stable over time. One potential caveat of this study is that the time span only covers 2 weeks, so the risk of respondents simply remembering their choices, that is, a carry-over effect, might be present. However, we received no indications from respondents that this was the case. Follow-up questions revealed that 68 per cent of the respondents did not remember what they had chosen 2 weeks ago in the original survey. Moreover, since we find no differences in correspondence rate between those stating to have remembered, we argue that a potential carry-over effect seems unlikely. Arguably, the benefit of keeping the time span relatively short is that the assumption of stable preferences is less likely to be violated due to external influences or shocks. Though, since we used an incentivised setting where respondents actually took home a bag of apples after each round of interviews, there was a small chance that respondents would be more familiar with the good in the retest since they had the chance of consuming the chosen apples at home. In case the respondent then tasted these apples and realised that she did not like them, it could be argued that such an experience would cause preferences to change in the retest. However, there was no indication from respondents that this was the case.

Obviously, it is not reasonable to generalise based on a single empirical data set, so further studies replicating our set-up and ideally using more respondents would seem relevant. Another issue for further research could be to assess the impact of using real incentives on the level of reliability, compared to purely hypothetical settings. Also, it would be interesting to investigate a similar within-sample, real incentives test–retest experimental design for a nonmarket good which could potentially yield very different results.

References

Alfnes, F., Guttormsen, A.G., Steine, G. and Kolstad, K. (2006). Consumers' willingness to pay for the color of salmon: a choice experiment with real economic incentives, *American Journal of Agricultural Economics* 88, 1050–1061.

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

Boman, M., Mattsson, L., Ericsson, G. and Kristrom, B. (2011). Moose hunting values in Sweden now and two decades ago: the Swedish hunters revisited, *Environmental and Resource Economics* 50, 515–530.

Brouwer, R. and Bateman, I. (2005). Temporal stability and transferability of models of willingness to pay for flood control and wetland conservation, *Water Resource Research* 41, 1–6.

Brownstone, D. and Train, K. (1999). Forecasting new product penetration with flexible substitution patterns, *Journal of Econometrics* 89, 109–129.

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

Carlsson, F. and Martinsson, P. (2001). Do hypothetical and actual marginal willingness to pay differ in choice experiments?, *Journal of Environmental Economics and Management* 41, 179–192.

Carlsson, F., Frykblom, P. and Lagerkvist, C.-J. (2005). Using cheap talk as a test of validity in choice experiments, *Economics Letters* 89, 147–152.

Carlsson, F., Mørkbak, M.R. and Olsen, S.B. (2012). The first time is the hardest: a test of ordering effects in choice experiments, *Journal of Choice Modelling* 5(2), 19–37.

Chang, J.B., Lusk, J.L. and Norwood, B. (2009). How closely do hypothetical surveys and laboratory experiments predict field behavior?, *American Journal of Agricultural Economics* 91(2), 518–534.

Cohen, J. (1968). Weighed kappa: nominal scale agreement with provision for scaled disagreement or partial credit, *Psychological Bulletin* 70(4), 213–220.

Cummings, R.G. and Taylor, L.O. (1999). Unbiased value estimates for environmental goods: a cheap talk design for the contingent valuation method, *American Economic Review* 89(3), 649–665.

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

Gravelle, H. and Rees, R. (1992). *Microeconomics*. Pearson Education Ltd, Harlow, UK.

Greene, W.H. (2003). *Econometric Analysis*. Prentice-Hall International Inc, New York.

Greene, W.H. and Hensher, D.A. (2007). Heteroscedastic control for random coefficients and error components in mixed logit, *Transportation Research Part E* 43(5), 610–623.

Hanley, N., Wright, R.E. and Koop, G. (2002). Modelling recreation demand using choice experiments: climbing in Scotland, *Environmental and Resource Economics* 22, 449–466.

Herriges, J.A. and Phaneuf, D. (2002). Inducing patterns of correlation and substitution in repeated nested logit models of recreation demand, *American Journal of Agricultural Economics* 84(4), 1076–1090.

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

Lancaster, K.J. (1966). A new approach to consumer theory, *Journal of Political Economy* 74 (2), 132–157.

Liebe, U., Meyerhoff, J. and Hartje, V. (2012). Test-retest reliability of choice experiments in environmental valuation, *Environmental and Resource Economics* 53(3), 389–407.

Liekens, I., Schaafsma, M., Brouwer, R. and de Nocker, L. (2012). *Temporal stability of preference and willingness to pay elicitation in choice experiments: a test-retest*. European Association of Environmental and Resource Economists 19th Annual Conference 27–30 June 2012, Prague.

List, J.A., Sinha, P. and Taylor, M.H. (2006). Using choice experiments to value non-market goods and services: evidence from field experiments, *Advanced Economic Analysis and Politics* 6(2), 1–37.

Luce, R.D. (1959). *Individual Choice Behaviour*. Wiley, New York.

Lusk, J.L. and Schroeder, T.C. (2004). Are choice experiments incentive compatible? a test with quality differentiated beef steaks, *American Journal of Agricultural Economics* 86(2), 467–482.

McConnell, K.E., Strand, I.E. and Valdés, S. (1998). Testing temporal reliability and carry-over effect: the role of correlated responses in test-retest reliability studies, *Environmental and Resource Economics* 12, 357–374.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior, in Zarembka, P. (ed), *Frontiers in Econometrics*. Academic, New York.

Miguel, F.S., Ryan, M. and Scott, A. (2002). Are preferences stable? The case of health care, *Journal of Economics Behavior and Organization* 48, 1–14.

Mitchell, R.C. and Carson, R.T. (1989). *Using Surveys to Value Public Goods: The Contingent Valuation Method*. Resources for the Future, Washington, DC, USA.

Olsen, S.B., Lundhede, T., Jacobsen, J.B. and Thorsen, B.J. (2011). Tough and easy choices: testing the influence of utility difference on stated certainty-in-choice in choice experiments, *Environmental and Resource Economics* 49, 491–510.

Rigby, D. and Burton, M. (2011). *Intertemporal choice consistency and the information sensitivity of welfare estimates in stated preference studies*. EAERE 18th annual conference, Rome. June 29 to July 2.

Rose, J.M., Collins, A.T., Bliemer, M.C. and Hensher, D.A. (2009). *Ngene 1.0 Stated Choice Experiment Design Software*, University of Sydney, Sydney, Australia.

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

Scarpa, R. and Rose, J.M. (2008). Design efficiency for non-market valuation with choice modelling: how to measure it, what to report and why, *Australian Journal of Agricultural and Resource Economics* 52, 253–282.

Scarpa, R., Ferrini, S. and Willis, K.G. (2005). Performance of error component models for status-quo effects in choice experiments, in Scarpa, R. and Alberini, A. (eds), *Applications of Simulation Methods in Environmental and Resource Economics*. Springer Publisher, Dordrecht, the Netherlands.

Scarpa, R., Willis, K. and Acutt, M. (2007). Valuing externalities from water supply: status-quo, choice complexity and individual random effects in panel kernel logit analysis of choice experiments, *Journal of Environmental Planning and Management* 50(4), 449–466.

Scarpa, R., Thiene, M. and Marangon, F. (2008). Using flexible taste distributions to value collective reputation for environmentally-friendly production methods, *Canadian Journal of Agricultural Economics* 56, 145–162.

Scarpa, R., Zanoli, R., Bruschi, V. and Naspetti, S. (2013). Inferred and stated attribute non-attendance in food choice experiments, *American Journal of Agricultural Economics* 95(1), 165–180.

Shiell, A. and Hawe, P. (2006). Test-retest reliability of willingness to pay, *European Journal of Health Economics* 7, 176–181.

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

Swait, J. and Louviere, J. (1993). The role of the scale parameter in the estimation and comparison of multinomial logit models, *Journal of Marketing Research* 30, 305–314.
Train, K. (2003). *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge.

Appendix

Table A1 Socio-demographic distribution of sample

	Variable definition	Mean	Std. Dev.
Gender	1 = man; 2 = woman	1.79	0.41
Age	Respondents age	50.50	17.38
# Children	Number of children living at home	0.68	0.93
HH income	Annual household income (in DKK 1000s)	450.00	325.65

Table A2 EC model results including LR test statistics

	Model (i) Split 1 – test		Model (ii) Split 2 – retest		Model (iii) Pooled without scale correction		Model (iv) Pooled with scale correction	
	Coefficient (Std. Err.)	*t*-value	Coefficient (Std. Err.)	*t*-value	Coefficient (Std. Err.)	*t*-value	Coefficient (Std. Err.)	*t*-value
ASC SQ	−3.21 (1.21)	2.64***	−2.33 (1.48)	1.57 ns	−2.51 (1.1)	2.29**	−0.855 (0.37)	2.31**
Organic produce	0.856 (0.132)	6.49***	0.959 (0.173)	5.54***	0.896 (0.112)	8.01***	0.3 (0.0543)	5.52***
Local produce within Denmark	1.26 (0.253)	5.00***	1.46 (0.417)	3.49***	1.34 (0.252)	5.31***	0.448 (0.0933)	4.81***
Danish produce	−0.69 (0.301)	2.29**	−0.98 (0.327)	2.99***	−0.819 (0.278)	2.94***	−0.277 (0.101)	2.73***
Green coloured apples	0.395 (0.234)	1.69*	0.894 (0.403)	2.22**	0.596 (0.21)	2.84***	0.206 (0.075)	2.75***
Yellow coloured apples	1.56 (0.375)	4.17***	2.43 (0.455)	5.33***	1.9 (0.292)	6.52***	0.648 (0.125)	5.2***
Red coloured apples	0.879 (0.359)	2.45**	1.83 (0.397)	4.6***	1.26 (0.262)	4.81***	0.435 (0.0992)	4.39***
Sweet and mealy apples	0.487 (0.277)	1.76*	1.1 (0.363)	3.02***	0.772 (0.26)	2.97***	0.265 (0.0925)	2.87***
Price	−0.0215 (0.0089)	2.42**	−0.0248 (0.0134)	1.84*	−0.0206 (0.00939)	2.2**	−0.00698 (0.00323)	2.16**
Error Component	2.76 (0.863)	3.19***	2.95 (0.621)	4.75***	2.64 (0.524)	5.05***	0.932 (0.198)	4.7***
Scale (retest; test = 1)							1.18 (0.224)	0.80 ns
N	300		300		600		600	
# Respondents	25		25		25		25	
LL	−178.165		−159.754		−346.171		−345.257	
Pseudo R^2	0.429		0.485		0.46		0.46	
LR test statistics					16.504		14.676	

Notes: ***, ** and * indicate the 1%, 5% and 10% levels of significance – ns indicates insignificance.