Revisiting preference stability: a first step to testing the incentive compatibility arguments in transport research

Romain Crastes dit Sourd[1], Stephane Hess[1], Maria Börjesson[2], David Hensher[3]

Address for correspondence

Romain Crastes dit Sourd

Institute for Transport Studies and Choice Modelling Centre, University of Leeds

UK-Leeds LS2 9JT

R.CrastesditSourd@leeds.ac.uk

[1]Institute for Transport Studies and Choice Modelling Centre, University of Leeds,

[2]Centre for Transport Studies, KTH,

[3]Institute of Transport and Logistics Studies, The University of Sydney Business School

Acknowledgements

The authors would like to thank Mikołaj Czajkowski for his valuable contributions to this work. The views expressed in this article are the authors' own. The authors acknowledge the financial support by the European Research Council through the 48 consolidator grant 615596-DECISIONS.

Abstract

The literature on *incentive compatibility* has explored whether stated preference survey designs encourage respondents to answer truthfully. The work of Carson and Groves (2007) has received a lot of attention from environmental economists, but is also raises issues of concern for transport analysts because it suggests that respondents act strategically after the first choice task. In this paper, we make use of choice task specific models to uncover the relations between the first choice task and the subsequent ones and we provide a series of tests to analyse trends within and across data from 8 surveys in transport and environment.

Final version: 16 January 2018

Keywords: incentive compatibility; stated choice; consistency; strategic behavior; value of time; willingness to pay

1.0. Introduction

Stated choice (SC) is a popular survey design for studying choice behaviour. This methodology has been used for several decades across many areas of research, especially transport, marketing, health and environmental and resource economics. A key application of SC data and the subsequent estimation of discrete choice models is the derivation of monetary valuations, commonly referred to as willingness-to-pay (WTP) measures. These look at the monetary value that respondents place on a unit change in the characteristics of products or alternatives, and can be either for private goods (e.g. travel time) or public goods (e.g. forest preservation). While much of the emphasis is on marginal WTP, work especially in environmental economics has also looked at the WTP for an entire 'package' or 'programme' (Crastes *et al*., 2014).

The use of SP techniques has become a widely accepted tool for producing policy guidance, but criticism has never abated completely. While all fields in which SP techniques are used have their strong believers as well as opponents of the technique, nowhere has the level of debate been as extensive as when applied to non-market valuation. Indeed, environmental economics mainly addresses the management of complex public goods, such as forest biodiversity, which has given rise to many concerns regarding whether respondents are familiar enough with the good being valued (LaRiviere *et al.*, 2014), are not acting strategically during the survey (Scheufele and Bennett, 2012) and believe in the survey settings overall. These concerns however are again not area specific, and transport modellers have similarly discussed hypothetical and strategic bias in great detail (Hensher, 2010; Loomis, 2014).

In environmental economics, following the seminal theoretical work of Carson and Groves (2007), a large stream of studies has attempted to define the survey conditions under which respondents provide reliable answers, which is known as *incentive compatibility* (Taylor *et al*., 2010; Scheufele and Bennett, 2012; Vossler *et al*, 2012). The conditions for a survey to be incentive compatible have been discussed by Mc Nair *et al*., (2011) and summarized by Czajkowski *et al*. (2015) as follows:

i) respondents should correctly understand and answer the question being asked (including the requirement that the good(s) being valued, including the different attribute levels and cost, are seen as plausible);

ii) respondents need to see the survey as consequential, that is their responses should be viewed as potentially influencing agency's actions and agents must care about the outcomes;

iii) the payment has to be coercive, that is the payment vehicle must be able to impose costs on all agents if the government undertakes the project and the survey should be seen as take-it-or-leave-it offer, so that agents do not see their decisions as influencing any other offers that may be made;

iv) the message space of a choice question cannot be larger than a single binary comparison without restricting the space of allowable preference functions, that is a single binary choice is the only elicitation format that has a potential to be incentive compatible.

Many users of SC surveys in transport would be unfamiliar with the above discussions in environmental economics. Moreover, conditions (ii) and (iii) may be less relevant in transport because respondents are usually not asked to evaluate policies but private goods, albeit that, as we will see, some of the same concerns have been raised. We will now discuss each of the four components in turn and contrast them with current practice in transport research. In doing so, we specifically take no position as to the theoretical validity of these conditions.

Condition (i) has received more attention from the environmental economics literature, which may be due to the fact that transport practitioners usually seek to elicit preferences for familiar goods such as journey characteristics (Hensher, 1994), as opposed to complex goods such as biodiversity (Bateman, 2015). However, unfamiliar topics also apply in transport, for example looking at choices involving new modes of transport (Li *et al.*, 2015) or choices involving goods that a respondent is unfamiliar with making trade-offs for, such as road safety measures (Rizzi and Ortúzar, 2003). Analysts would use the term hypothetical bias to refer to this issue. In order to make sure that respondents correctly understand and answer the questions being asked, researchers have experimented with several approaches. Examples include the framing of questions, the provision of a priori information or the use of cheap-talk (the respondent must "think-out-loud" while choosing the alternative they prefer) or the oath (the respondent must sign a paper where they promise to answer truthfully).

Condition (ii) would be fulfilled if the following conditions are met:

a. *Consequentiality*, respondents must believe that policy makers will *actually* consider their preferences and consent to not deliberately misrepresent their preferences in order to influence the decision making process (Hanley, Mourato and Wright, 2001).
b. *No strategic bias*: assuming that a survey is consequential does not guarantee that answers will be truthful. Respondents may be involuntarily incentivised to misrepresent/misreport their preferences for their own benefit.

In non-market valuation, *consequentiality* is usually introduced by the means of a *consequentiality script*. More precisely, respondents must read a short text designed to increase respondents' perception that their participation will have influence on an important outcome. This is very different from the approach used in transport research, where only strategic bias is a specific concern, and the solutions to mitigate it are radically different. Indeed, taking for example a toll road setting in a survey, a respondent may understate their WTP to prevent the road from being built, while, when built, they might actually use it. For this reason, SC surveys in transport often explicitly mask the aim behind the study or hide the policy implications (of course, this does not apply to all applications). This is the opposite in non-market valuation where the policy implications must be absolutely clear otherwise consequentiality is lost.

Condition (iii) has received less attention from the literature, although efforts have begun to explore the use of follow-up questions to measure whether respondents believe that the decision makers will be able to impose costs on all agents according to Czajkowski *et al*. (2015).

Conditions (iii) may be too restrictive for transport practitioners. Indeed, for example a toll road is a club good (unless it is congested, in which case it becomes a private good) and the government cannot force individual agents to buy private goods or club goods so such survey setting should be considered as unrealistic.

Condition (iv) first suggests that respondents should not face more than two alternatives according to the theoretical model proposed by Carson and Groves (2007). However, Hensher (2006a) as well as Meyerhoff and Liebe (2009) and Meyerhoff *et* al. (2013) have investigated the effects of the number of alternatives on preferences and found that the number of alternatives only has a marginal effect on preferences. More particularly, (Hensher, 2006b) found that differences in behavioural responses due to the dimensionality of SC surveys exist when each dimension is assessed without controlling for the dimensions. Second, condition (iv) suggests that repeated choice tasks are not *incentive compatible* and respondents should only face a single choice task. Indeed, since respondents may be made aware in advance of having to face multiple choice tasks, they can exploit information about previous choice tasks and decisions as they go through the survey and act strategically rather than revealing their true preferences (Czajkowski *et al*., 2014). This requirement would invalidate essentially all SC work in transport.

It is the belief of the authors of this paper that there is an urgent need for a critical appraisal of the validity of the arguments on incentive compatibility for transport research. Indeed, it is clear that very few SC surveys are actually *incentive compatible* according to the conditions mentioned above. We believe that there is general agreement on condition (i), that is ensuring that results are not unduly affect by the hypothetical setting of the choice scenarios. Condition (ii), which requires that respondents see the survey as consequential requires a philosophical discussion, contrasting the conflicting aims of avoiding strategic bias while ensuring valid choices. Condition (iii) seems to be less relevant in a transport context, and can possibly be largely ignored.

Our focus in this paper is on condition (iv) as it is the one that would invalidate most existing studies. We believe that the work in the design of designs studies (Hensher 2006c) has already addressed the issue of the number of alternatives in a choice task and provided no evidence to suggest a requirement to stick to binary choices. We address the second component of condition (iv), namely the requirements for just a single choice task rather than a repeated setting.

Scheufele and Bennett (2012), among several other authors, have empirically proven that repeated choice tasks may induce strategic behaviour in non-market valuation surveys. The authors suggest that, as by design, multiple successive choice sets in a given survey may feature similar (or identical) cost values for different levels of provision of a particular good, this may cause respondents to question the credibility of the survey. Ultimately, this loss of credibility may drive respondents to reject a preferred choice option when the same or a similar option was offered in a previous choice question at a lower price. Scheufele and Bennett (2012) conclude their particularly important contribution with a dilemma: if respondents become aware of and take advantage of strategic opportunities during the survey, then one may conclude that the first choice task better reflects respondents' true preferences. However, if respondents learn about their preferences during the survey and adjust their answers after the first choice task, then the authors conclude that the first answer should be excluded from the analysis. Czajkowski *et* al. (2014) as well as Meyerhoff and Glenk (2015), among others, found strong evidences of learning effects in the literature. Moreover, Börjesson and Algers (2011) and Börjesson and Fosgerau (2015) found that response time declines over the sequence of choices, which is also an additional proof that respondents learn about both their preferences and the survey settings as they progress through the survey.

Several recent studies (some of which are still under review) have investigated how to make SC surveys *incentive compatible* while allowing for repeated choices. As an illustration, (Mahieu *et al.,* 2015) have introduced a lie detection device in a survey on tree planting. The participants were told that they would not receive payment for their participation in the survey and their answers would not be considered for policy making if they provide untruthful answers. The results indicate that the variance of the error term decreases only when respondents fully engage in the survey by believing in the lie detection device. These attempts to reinforce the incentive compatibility of the repeated SC survey format should not cover the fact that the non-market valuation literature concludes that it remains unknown whether repeated choice experiments are incentive compatible. The key question in this context is whether the same behaviour drives choices in all choice tasks (Czajkowski *et al*., 2015)?

The aim of this paper is to address this question by investigating preference stability across choice tasks. Our work shares some similarities in the approach and data used by Hess *et al*., (2012) when looking at fatigue effects, but our focus is different. We develop the idea that *incentive compatibility* could be related to the anchoring effect, which is the one of the most well-established cognitive heuristics in behavioural economics (Tversky and Kahneman, 1974). Translated to an SP setting an anchor effect would imply that the attributes in the first choice would set an anchor in the respondents' minds, and then be used as mental reference points in the following sequence of choices. It is different from *incentive compatibility* as previously defined because it does not necessarily imply that the valuation uncovered by the first choice would be lower or higher, but that the attribute levels of the first choice would influence the preferences elicited from the rest of the choice sequence.

We look at a number of datasets collected in various countries and investigate the extent of any difference between the responses provided for the first choice task and the subsequent ones. We use data from 8 surveys, covering both transport and non-market valuation to establish whether the context leads to differences. We also break the correlation between attributes of the tasks and choice behaviour by only using datasets where the order of choice tasks has been randomized across respondents.

The remainder of this paper is organised as follows. In the next section, we outline the empirical testing framework used in our analysis. In section 3, we present the different datasets used in our empirical work. Section 4 presents model results and finally, Section 5 concludes.

2.0. Framework for empirical tests

This section gives a brief outline of the framework used for the empirical tests conducted for this study. As previously introduced, our empirical testing frameworks makes use of choice task specific models. More precisely, we estimate a separate model for each choice task. We estimate choice task specific Multinomial Logit models (MNL) in order to find (i) whether the responses to the first choice task are truly different from the subsequent ones, (ii) whether a specific response pattern can be identified across choice tasks and (iii) whether this pattern is similar between transport and non-market valuation surveys. We compare the results within each dataset using t-ratio tests and we compare our results across datasets using graphs analysis and k-density tests. About 150 models have been estimated in total in order to ensure the robustness of our results. We restrict our analysis to MNL because of sample size limitations. In addition, we also estimate:

(i) A complete model, that is a model estimated using all the choice tasks for a given dataset.

(ii) A restricted model, that is a model estimated without using the first choice task.

The model results are then analysed by the mean of a series of statistical tests, as follows.

2.1. Test 1: Graphs analysis

Our first test simply consists in mapping the coefficient values of each of the choice task specific models and to then compare the trends both within and across datasets. In doing this, we restrict our analysis to one environment dataset and one transport dataset for clarity purpose. However, we provide all the other model results so that the reader can replicate our approach if needed.

2.2. Test 2: T-tests

The second test is a T-test that seeks to identify, for each dataset, whether the parameters of the first choice task specific model are significantly different than the parameters of the other choice tasks specific models. More precisely, for each dataset and each parameter, we estimate a T-test between the coefficient value for the first choice task specific model and each of the other choice task specific model. A significant value means than a different behavior drives choices during the first choice task, which suggests that a repeated SC survey design may not be fully *incentive compatible*.

2.3. Test 3: Trends in relative importance of different attributes in comparison to the first choice task

The third test seeks to identify similarities within and across datasets in how the values of the parameters differ across choice tasks. We make use of the results from the choice tasks specific models. We calculate $R_{dpt} = \beta_{dp1}/\beta_{dpt}$. β_{p1} is the value of parameter p in the first choice task specific model and β_{pt} is the value of the same parameter for the *t*th choice task specific model. *d* corresponds to a given dataset. For example, the Atlanta dataset presented below has 8 choice tasks and 5 parameters so we will calculate $R_{p1} = \frac{\beta_{p1}}{\beta_{p2}}, R_{p2} = \frac{\beta_{p1}}{\beta_{p3}}, \dots, R_{p7} = \beta_{p1}/\beta_{p8}$ for each of the 5 parameters. We then plot the kernel density estimate of each of the ratio distributions. Kernel density estimation is simply a non-parametric technique for estimating the probability density function (PDF) of a random variable. Finally, the k-density test for comparing the common area of kernel density estimates proposed by Martínez-Camblor et al. (2008) and introduced in the field of non-market valuation by Crastes et al. (2015) is used. This test allows us to assess whether there are similarities in how the values of the parameter differ across choice tasks. More precisely, the k-density test gives a simple measure of the proximity of two kernel density estimates. This measure, known as the *AC* statistic, is comprised between 0 and 1. A value of 0 corresponds to an absolute discordance while a value of 1 corresponds to an absolute match of the distributions. The test is illustrated for three arbitrary kernel densities f, g and h by Figure 1 below where the grey area gives a measure of the *AC* statistic.

[Figure 1 here]

In the context of this paper, a value above 0.5 means that two ratio distributions tend to be similar (Martínez-Camblor *et al*., 2008), which suggests similarities in how the values of the parameter differ across choice tasks. This test can be used to compare differences across datasets. In this paper, we apply this test between each of the pairs of coefficient ratio PDF for two of the datasets featured in this paper. We chose similar datasets (in terms of number of attributes and number of choice tasks), one dataset from transport and one dataset from environment, in order to test not just whether there are similarities across parameters in individual datasets but whether similarities exist also across datasets. Moreover, we run the same tests for all of the coefficient ratio PDF in order to identify whether there are similarities in how the values of the parameters differ across choice tasks overall.

3.0. Data

In this section, we present the different transport and environment datasets used in our analysis. These datasets were collected in different countries (Australia, Denmark, Poland, UK, USA) and vary in terms of design (number of attributes, number of choice scenarios, and number of alternatives). Overall, we use two datasets from non-market valuation surveys, six datasets from transport surveys and one dataset which investigates the WTP for funding theatres in Warsaw (Poland). By including data from such a diverse set of surveys, we can establish whether differences exist across areas and whether this justifies the relative lack of concern by transport analysts on this matter. As previously stated, the common factor across all datasets is that for each survey participant, the order of the choice tasks was randomised.

3.1. Transport data

3.11. Atlanta toll road study

The first study used data collected in 2008 and presented in Hess *et al*., (2008). For each respondent, data was collected from eight choice tasks (3 alternatives per choice task: driving in the existing untolled lanes, driving in a tolled lane or carpooling in the managed lane in return for a reduced toll). 4 different treatments using different experimental design were applied. As for the rest of the transport datasets used in this paper, further details can also be found in Hess *et al*. (2012).

3.12. First Australian toll road dataset

Our second case study makes use of data from a three alternative route choice experiment in Australia (one alternative consisted in a reference trip and was kept fixed across choice tasks). The alternatives were described in term of free flow time (*ff*), slowed down time *(sdt*), running costs (*costs*), tolls (*tolls*) and travel time variability (*var*). More details can be found in Hensher and Rose (2005).

3.13. Second Australian dataset

The Second Australian dataset is very similar to the First Australian toll road study. It makes use of an additional travel time component, crawl time (*crawl*). Each respondent was faced with 16 choice tasks. See Hess *et al*. (2010) for more details.

3.14. Danish value of time data

Our fourth case study makes use of data from a choice experiment conducted in Denmark (more details can be found in Fosgerau (2006). In each of the 8 choice tasks, a respondent was faced with two unlabelled alternatives described by travel time (*tt*) and travel cost (*tc*).

3.15. Fungibility data

The fifth case study used data looking at the relative sensitivities to rail travel time, cost and safety, defined in terms of number of accidents. Each respondent faced three different binary stated choice experiment which consisted in trading time against cost, time against safety and safety against cost. 15 choice tasks were used in total. More details can be found in Orr *et al*. (2010).

3.2. Non-market valuation data

3.21. - Ecological value of Polish forests survey

The sixth case study used data collected by Czajkowski *et al*. (2014). The survey consisted in eliciting the preferences of the general public in Poland for increasing the amount of recreational infrastructures in Polish forests and protecting biodiversity. Respondent faced 26 choice tasks made of four alternatives. Each alternative was described with 4 attributes taking several levels *nat1* and *nat*2, *tra*1 and *tra*2, *inf*1 and *inf* 2, all related to partial and substantial increase of forest quality. A price attribute was also included.

3.22. Białowieża Forest survey

The seventh case study used data collected by Bartczak (2015). The survey features 12 choices tasks consisting of 3 alternatives each and uses attributes describing changes in the quality of the Białowieża forest (Poland). The choice experiment comprised 4 attributes: *cen* (level of naturalness of the commercial part of the forest, *gos* (level of naturalness of the second-growth forest), *vis1* and *vis2* (restrictions on number of visitors per day) and *fee* (annual cost per household).

3.23. Warsaw theatres survey

The eight case study used data from a repeated choice experiments on public support for culture (Czajkowski *et al*., 2015). More precisely, the survey relates to the introduction a program of highly discounted tickets (12 choice tasks per respondents). In each choice task, respondents were asked to choose one of the two alternatives presented, including one *status quo* alternative. Each alternative was composed of 4 binary attributes corresponding to the type of play being funded: *roz* (entertainment theatre), *sro* (drama theatre), *dzi* (children's theatre) and *eks* (experimental theatre). A cost attribute was also included.

4.0. Empirical analysis

In this section, we discuss the empirical results. The model specification for each of the eight datasets makes use of alternative specific constant (ASC) for each of the alternatives except one, while all the models are MNL. It is worth noting that we use data from both labelled and unlabeled SC surveys so the meaning of the different ASCs must be carefully interpreted. We look at the choice task specific model results and also the results of the various tests in turn for each of the datasets.

The choice task specific MNL results for the transport surveys are summarized in Table A1 to A5 while the results for the non market valuation surveys are summarized in Table A6 to A8 (see online appendix[1]). We first look at trends in model fit. For each of the models, the model fit of the first choice task specific model does not reveal any specific trend. More precisely, the first choice task specific model for the Polish forest data obtained a log-likelihood of -1191.07 which is the second worst fit for this dataset while the fit for the first task is the second best result for the Białowieża Forest data (-1221.01) and the worst fit for the Warsaw theatre data (-1004.61). The transport survey results also show a remarkable absence of pattern: The first choice task specific models for the 3rd and 5th stage of the Atlanta survey obtained the worst fit (respectively -791.08 and -1540.52) while the results for the first Australian dataset are mixed and do not allow to detect any specific trend. The MNL model for the Second Australian dataset obtained the second worst fit (-194.49) while the results for the Danish surveys exhibited the best fit (-798.38

[1] The complete model results are available in the online appendix: https://goo.gl/5UKBVk

for the non-commuters and -215.97 for the commuters respectively). Finally the results for the first choice task specific model for the fungibility data exhibited the worst fit (-241.62). These findings provide some evidence that the choices made by respondents in the first task are *more random* than those in later tasks, but this is not a universal finding. These results alone only indicate that choices for the first choice task *may not* be driven by the same behaviour than the rest of the choices in some cases. However, it also shows that there is no clear pattern. As a result, we turn our attention next to the first test and study the impact that estimating choice task specific models has on model parameters.

We obtain mixed results. The first non-market valuation dataset (Polish forests) reveals that three attributes parameters (*nat1*, *nat2* and *tra1)* are the lowest for the first choice task specific model, while the Białowieża Forest dataset results do not exhibit such behaviour (several parameters are not significant) and the Warsaw dataset results only report that *sro* is the smallest attribute parameter for the first choice task specific model. The results for the transport datasets also show no evidence of pattern or trend. Apart from the Danish data results, which systematically exhibit smaller attribute parameter estimates for the first choice task specific model and higher ASC values. The results for the first Australian dataset do not reveal any specific trend (some coefficients are smaller for the first choice task specific models but they remain overall very close to the results of the models estimated for the subsequent choice tasks). For the second Australian dataset, *tt* and *var* are found to be the lowest for the first choice set specific model. The Danish data only report very high ASC for the first model (for both commuters and non-commuters). Finally, no significant trend can be derived from the fungibility data.

Figure 2 and 3 provide a graphical representation of the pattern followed by the choice task specific model coefficients. We have chosen to represent and compare the results derived from one non-market valuation dataset (the Białowieża Forest survey) and one transport dataset (the second Australian survey). Both datasets are similar in terms of number of attributes, number of choice tasks and number of alternatives. Figure 2 and 3 explicitly show that in a vast majority of cases, the parameters follow a stable pattern which seems mostly due to natural variations rather than a strategic reaction generated after the first choice task, in which case we would observe that the parameter values tend to follow a specific direction after the first choice task. For the first dataset, the attribute parameters for *vis1* and *fee* go upward after the first choice task. We have not identified such pattern for the second Australian survey. Altogether, these results suggest that respondents do not significantly change preferences while progressing across choice tasks. Moreover, the fact that a clear response pattern cannot be identified suggests that the respondents who supposedly act strategically do not have a strong influence on results. Tests 2 and 3 below confirm these findings.

[Figure 2 and 3 here]

Results for the second test (T-test) can be found in Tables 1 to 4. Significant differences between parameter values appear in bold. Again, the second test does not allow to firmly conclude that a different behavior drives the first choice task. Overall, we find that the first choice task specific models exhibit significant differences in parameter values with the other choice task specific models in only 30 per cent of the cases. It is worth noting that these results are very heterogeneous across datasets. T-tests indicate that preferences do not vary at all between the first choice task specific model and the rests of the choice task specific models for the Białowieża Forest data. The parameters are different in 2.6 per cent of the cases for the Warsaw theatres dataset, 3.8 per cent for the second Australian dataset and 4.7 per cent for the fungibility dataset. On the other hand, the parameters are different in 42.8 per cent of the cases for the Danish dataset (for both commuters and non-commuters) and up to 62.5 per cent for the ecological value of Polish forests datasets. We do not find differences between non-market valuation dataset and transport research datasets.

[Tables 1 to 4 here]

Finally, we analyse the results from the k-density test for comparing the common area of kernel density estimates of the individual model coefficient values relative to the first choice task specific model. Test results can be found in Table 10. The test results reveal that there is a weak concordance between densities (that is $0.5<AC<0.75$) in only 29 cases out of 105 (about 27 per cent of the cases). However, we have not found any significant strong concordance between densities ($AC > 0.75$), which suggests that the relative difference of the parameter values estimated for the first choice task specific model and the subsequent choice tasks does not follow a common pattern, both within and across datasets. In addition, we computed the *AC* value for all of the kernel density estimates of individual model coefficient values relative to the first choice task specific model, for each dataset and each parameter. We found a very significant common area of 0.007 (significant at the 1 per cent level), which means that there is no concordance across datasets overall, that is no pattern[2].

[Table 5 here]

5.0. Conclusion

This is an attempt to provide a reliable although limited answer to of the ongoing debate on whether the repeated SC survey format, which is by far the most widely used format in transport research, has negative consequences on *incentive* compatibility. Indeed, it is supposed that respondent may adopt different behaviours as they move across choice tasks. Carson and Groves (2007) suggest that respondents may act strategically or randomly after they have completed the first choice task. We based our analysis on 8 datasets from both environment and transport in order to determine whether differences exist across fields and used a series of tests to compare results both within and across datasets. More precisely, we analysed the consistencies of choice

[2] Detailed results are available from the authors upon request.

task specific model results as the respondents move across choice tasks. We have formulated the hypothesis that if a survey is not incentive compatible, then we should be able to observe either a direction across choice task specific coefficients (that is the respondents chose to act strategically) or a completely erratic pattern (that is the models report completely inconsistent preferences).

Our findings show that the choice task specific models provide consistent estimates in a vast majority of cases. Little evidence has been found for supporting the presence of strategic behaviour when using repeated choice experiments. The use of several statistical tests including t-tests and k-density tests for measuring the common area of kernel density estimates show that respondents do not necessarily exhibit a specific behaviour during the first choice task in comparison with subsequent ones and also that there is no significant pattern in differences between coefficient values across choice task specific models. One argument against the approach we propose may be that *incentive compatibility* should be considered as a whole and that it may not be possible to say something about one of the conditions defining *incentive compatibility* if some of the others are not satisfied. For example, the first choice task is not necessarily *incentive compatible* if more than two alternatives are used or if the respondents do not know exactly the objective eof the survey, etc. Moreover, one could argue that transport surveys are not necessarily *incentive compatible*, for example when they address quasi-public goods (only an individual who chooses to use a new road pays for using it) so the payment is not consequential (respondents always have incentives to say that they would pay, only to have the good provided and be able to choose whether they want to use the road or not, once they see the final price). As previously discussed, we believe that there exist fundamental differences between transport and environmental research and there should be a general agreement across research fields about most of the conditions necessary for i*ncentive compatibility* but there is also an urgent need for a critical appraisal of the validity of some of the arguments on *incentive* *compatibility* for SC practitioners outside of environmental research.

The evidence in the present paper should contribute to fuel the debate about whether a repeated SC survey design is truly not *incentive compatible*. However, *incentive compatibility* is a broad topic and our results are limited to analyse whether respondents express the same behavior in the first choice task and in the rest of the SC survey. Although our results support the idea that the choice responses given by the respondents after the first choice task may not be as biased as what is suggested by Carson and Groves and by the literature on incentive compatibility in general, more research on this topic, especially comparisons across research fields, are needed. While we have not found any conclusive evidence to support the concerns in the non-market valuation literature, we call for a deeper engagement by SC practitioners with this topic.

References

1. Bartczak, A. (2015): 'The role of social and environmental attitudes in non-market valuation: An application to the Białowieża Forest', *Forest Policy and Economics,* 50, 357-365.

2. Bateman, A. J., Coobes, E., Fitzherbert, E., Binner, A., Bad'ura, T., Carbone, C., Fisher, B., Naidoo, R. and Watkinson, A.R. (2015): 'Conserving tropical biodiversity via market forces and spatial targeting', *Proceedings of the National Academy of Sciences,* 112, 7408-7413.

3. Börjesson, M. and Algers, S. (2011): 'Properties of internet and telephone data collection methods in a stated choice value of time study context', *Journal of Choice Modelling,* 4, 1-19.

4. Börjesson, M. and Fosgerau, M. (2015): 'Response time patterns in a stated choice experiment', *Journal of Choice Modelling,* 14, 48-58.

5. Carson, R.T., and Groves, T. (2007): Incentive and informational properties of preference questions, *Environmental and resource economics,* 37, 181-210.

6. Crastes, R., Beaumais, O., Arkoun, O., Laroutis, D., Mahieu, P.A., Rulleau, B., Hassani-Taibi, S., Barbu, V.S. and Gaillard, D. (2014): 'Erosive runoff events in the European Union: Using

discrete choice experiment to assess the benefits of integrated management policies when preferences are heterogeneous', *Ecological Economics*, 102, 105-112

7. Czajkowski, M, Vossler, C.A., Budziński, W., Wiśniewska, A. and Zawojska, E. (2015): 'Addressing empirical challenges related to the incentive compatibility of stated preference methods', *Working Papers, Faculty of Economic Sciences, University of Warsaw*.
8. Czajkowski, M., Hanley, N., LaRiviere, J. and Simpson, K. (2014): 'What is the Causal Effect of Information and Learning about a Public Good on Willingness to Pay?' In *SIRE Discussion Papers*.
9. Czajkowski, M, M. Giergiczny and Greene, W. (2014): 'Learning and fatigue effect revisited – Investigating the effects of accounting for unobservable preference and scale heterogeneity', *Land Economics*, 90, 324 – 351.
10. Fosgerau, M. (2006): 'Investigating the distribution of the value of travel time savings', *Transportation Research Part B*, 40, 688-707.
11. Hanley, N., Mourato, S. and Wright, R.E. (2001): 'Choice modelling approaches: a superior alternative for environmental valuation?', *Journal of Economic Surveys*, 15, 435-462.
12. Hensher, D.A. (1994): 'Stated preference analysis of travel choices: the state of practice', *Transportation*, 21, 107-133.
13. Hensher, D.A. and Rose, J. (2005): 'Respondent Behavior in Discrete Choice Modeling with a Focus on the Valuation of Travel Time Savings', *Journal of Transportation and Statistics*, 8, 17-30.
14. Hensher, D.A., (2006a): 'Towards a method to establish comparable values of travel time savings from stated choice experiments with differing design dimensions'. *Transportation Research Part A*, 40, 829-840.
15. Hensher, D.A., (2006b): 'Revealing differences in behavioural response due to the dimensionality of stated choice designs: an initial assessment', *Environmental and Resource Economics*, 34, 7-44.

16. Hensher, D.A., (2006c): ‘How do Respondents Process Stated Choice Experiments? – Attribute consideration under varying information load’, *Journal of Applied Econometrics*, 21, 861-878.

17. Hensher, D.A., (2010): ‘Hypothetical bias, choice experiments and willingness to pay’, *Transportation Research Part B: Methodological*, 44, 735-752

18. Hess, S., D. Hensher and Daly, A. (2012): ‘Not bored yet - revisiting respondent fatigue in stated choice experiment’, *Transportation Research Part A: Policy and Practice*, 46, 626 - 644.

19. Hess, S., Smith, C., Falzarano, S. and Stubits, J. (2008): ‘Measuring the effects of different experimental designs and survey administration methods using an Atlanta Managed Lanes Stated Preference survey’, *Transportation Research Record*, 2049, 144-152.

20. Hess, S., Rose, J.M. and Polak, J.W. (2010): ‘Non-trading, lexicographic and inconsistent behaviour in stated choice data’, *Transportation Research Part D*, 15, 405-417.

21. Mahieu, P.A., Crastes dit Sourd, R., Louviere, J. and Zawojska, E. (2015) : ‘Lie detection in stated preference studies’. *Presented at the 3rd Workshop on non-market-valuation*, Nancy.

22. Martínez-Camblor, P., J. De Uña-Álvarez and Corral, N. (2008): ‘K-Sample test based on the common area of kernel density estimators’, *Journal of Statistical Planning and Inference*, 138, 4006-4020.

23. McNair B Bennett J and Hensher DA (2011) A comparison of responses to single and repeated discrete choice questions. In Resource and Energy Economics, Vol. 33, pp. 544-571.

24. LaRiviere, J., Czajkowski, M., Hanley, N., Aanesen, M., Falk-Petersen, J. and Tinch, D. (2014): ‘The value of familiarity: effects of knowledge and objective signals on willingness to pay for a public good’, *Journal of Environmental Economics and management*, 68, 376-389.

25. Li, Y., Schmocker, J., and Fujii, S. (2015): ‘Demand Adaptation Towards New Transport Modes: Case of High Speed Rail in Taiwan’, *Transportmetrica B: Transport Dynamics*, 3, 27-43.

26. Loomis, J. (2014): ‘Strategies for overcoming hypothetical bias in stated preference survey’, *Journal of Agricultural and Resource Economics*, 39, 34-46.

27. Meyerhoff, J. and Liebe, U. (2009): 'Status Quo Effect in Choice Experiments: Empirical Evidence on Attitudes and Choice Task Complexity', *Land Economics*, 85, 515–528.

28. Meyerhoff J and Glenk K (2015) Learning how to choose – effects of instructional choice sets in discrete choice experiments. In *Resource and Energy Economics*, Vol. 41, pp. 122-142.

29. Meyerhoff, J., Oehlmann, M. and Weller, P. (2015): 'The Influence of Design Dimensions on Stated Choices in an Environmental Context', *Environmental and Resource Economics*, 61, 385-407.

30. Orr, S., Hess, S. and Sheldon, R. (2010): 'Fungibility of monetary valuations in a transport context: an empirical investigation of the transferability of willingness to pay measures'. Presented at the European Transport Conference, Glasgow.

31. Rizzi, *L. and* Ortuzar J.D.D. (2003): 'Stated preference in the valuation of interurban road safety', *Accident Analysis and Prevention*, 35, 9-22.

32. Scheufele, G. and J. Bennett. (2012): 'Response strategies and learning in discrete choice experiments', *Environmental and Resource Economics*, 52, 435 – 453.

33. Taylor, L.O., Morrison, M.D. and Boyle, K.J. (2010): 'Exchange rules and the incentive compatibility of choice experiments', *Environmental and Resource Economics*, 47, 197-220.

34. Tversky, A. and Kahneman, D. (1974): 'Judgment under uncertainty: heuristics and biases', *Science*, New Series, 185, 1124-1131.

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

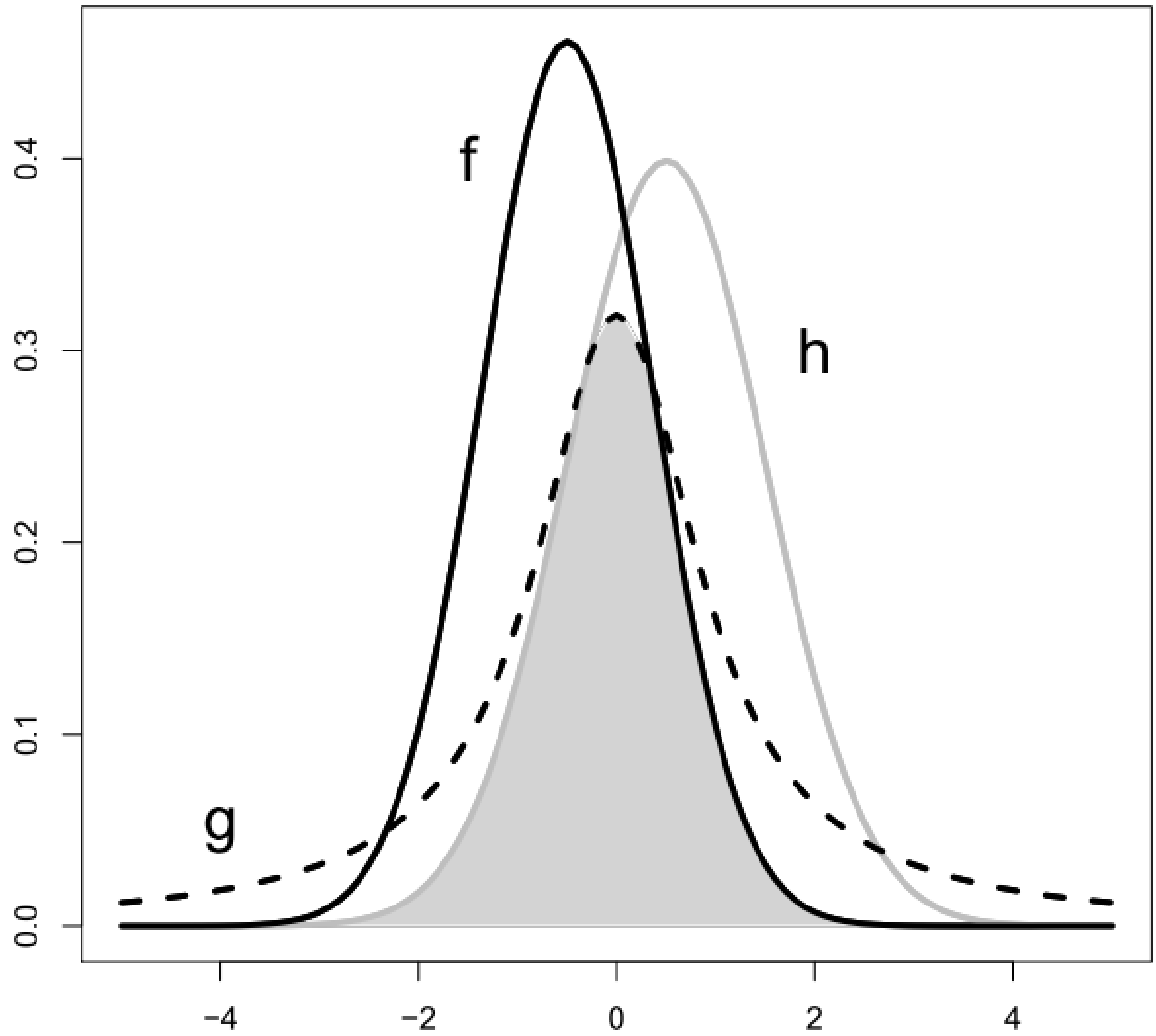

Figure 1 The *AC* test

Figure 2 Second Australian dataset: consistency across choice tasks

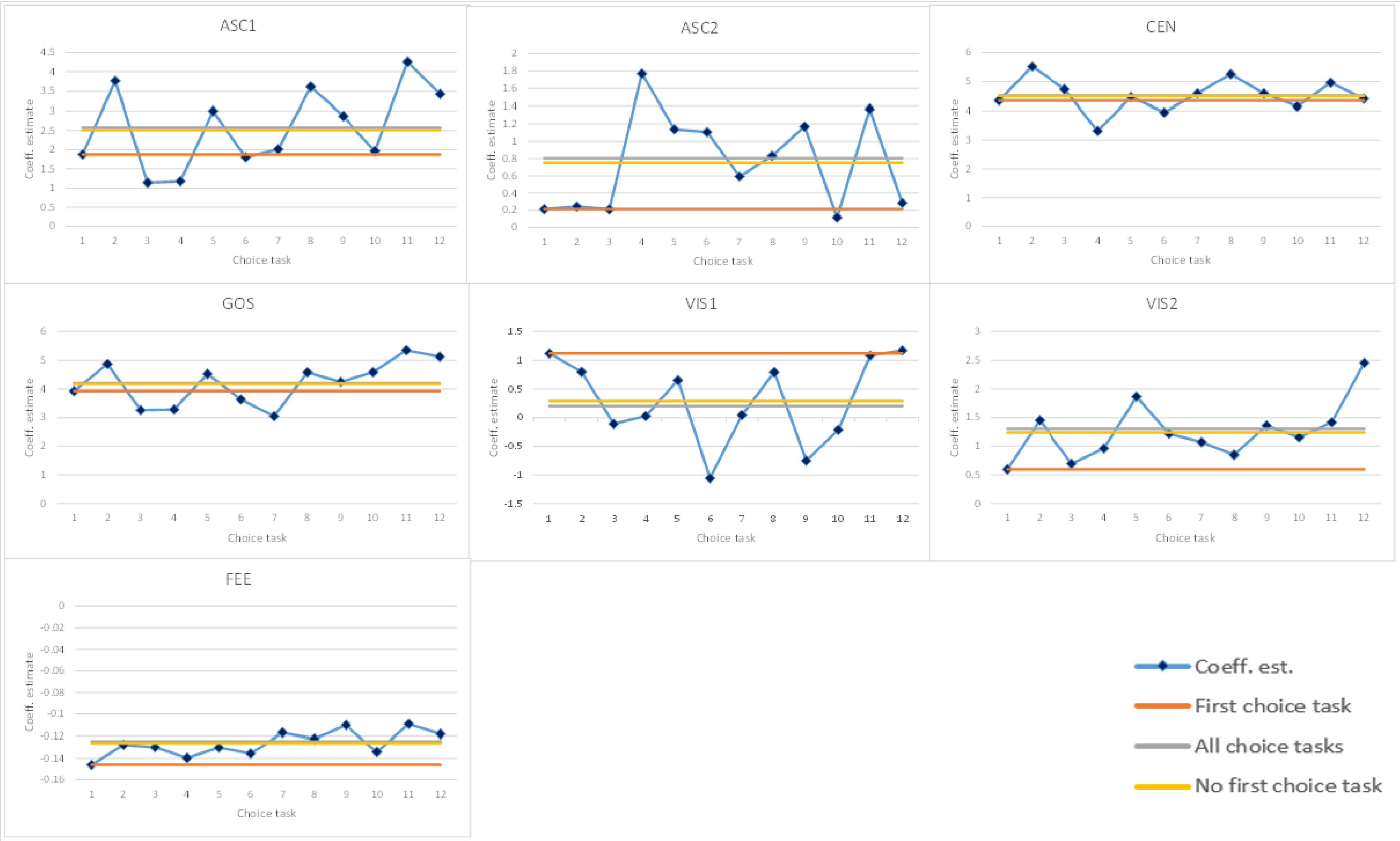

Figure 3 Białowieża forest data: consistency across choice tasks

TABLES

TABLE 1 T-tests

First Australian dataset								
	asc_rp	*asc_sp_2*	*cost*	*ff*	*sdt*	*Toll*	*var*	
Commuters								
2	-0.116	1.917	-0.287	-0.322	**-5.958**	0.796	1.175	
3	-0.936	0.404	-0.779	0.596	-0.678	1.049	1.262	
4	0.441	-0.102	-1.122	-0.672	0.151	-0.571	0.355	
5	-0.788	0.188	-0.702	0.909	-0.207	0.834	1.454	
6	-1.142	-0.112	0.495	-0.228	0.493	1.310	**2.290**	
7	-1.612	-0.853	0.355	0.673	1.267	0.765	1.200	
8	-1.176	0.366	-0.222	0.974	-0.375	1.793	1.573	
9	0.196	1.663	-0.669	0.458	**-3.629**	0.849	1.034	
10	-1.427	0.065	-0.221	1.848	-0.035	**2.407**	1.481	
11	-1.323	0.287	-0.223	1.547	-0.186	1.704	1.530	
12	-0.710	-1.008	0.000	0.240	**4.205**	1.580	0.772	
13	-0.945	-0.026	-0.146	0.841	0.030	1.795	**2.290**	
14	-1.037	-0.048	-0.072	0.738	0.065	1.711	1.752	
15	-1.709	-0.204	-1.011	**1.978**	0.103	**2.453**	1.528	
16	-0.645	-0.075	0.500	-0.332	0.226	1.470	0.999	
Non-commuters								
2	-0.453	0.391	1.181	1.786	-0.525	0.865	0.779	
3	-0.729	0.369	-0.207	-0.762	-0.259	0.000	0.900	
4	-1.915	-0.295	**-2.266**	**2.130**	0.368	-1.690	**2.364**	
5	**-2.149**	-0.430	1.120	1.364	-0.281	1.856	**2.520**	
6	-1.395	1.004	0.000	0.176	-1.144	0.111	0.992	
7	-1.538	0.571	-0.884	0.694	0.137	1.165	**2.137**	
8	-1.853	0.852	0.062	1.705	0.900	0.557	**2.303**	
9	-1.656	0.605	0.598	1.923	-1.563	0.111	**2.001**	
10	-1.070	0.789	0.543	0.925	-0.796	1.500	1.686	
11	-1.266	0.395	-1.210	1.750	0.374	0.561	1.419	
12	-1.357	0.934	-0.963	**-5.216**	**-5.182**	0.651	**2.165**	
13	-0.843	1.654	-0.535	1.166	-0.513	0.114	1.467	
14	-1.475	0.393	-0.752	0.390	-0.291	1.551	**2.027**	
15	-1.561	-0.391	-0.354	0.967	-0.365	-0.111	1.762	
16	-1.705	0.675	0.297	1.824	-0.820	0.000	1.547	
Second Australian dataset								
	asc_rp	*asc_sp_2*	*cost*	*ct*	*ff*	*sdt*	*toll*	*var*
2	-1.003	0.238	-1.291	-0.962	**-2.230**	0.527	-0.447	-1.506
3	-0.840	-0.580	-0.578	-1.637	-1.313	0.923	0.363	-1.314
4	-0.357	0.269	-0.879	1.440	-0.560	-0.353	-0.521	-0.552
5	-0.697	-0.213	-0.893	-0.608	-1.511	0.189	0.836	-1.407
6	-0.308	0.562	-0.598	-0.298	-0.551	1.164	0.302	-1.410
7	-0.251	-0.145	-0.604	-0.154	-0.331	1.765	0.296	-0.793
8	-1.840	-1.884	-1.892	-0.683	-1.872	0.087	0.074	-1.219
9	-1.529	-0.900	0.517	-0.232	-1.777	0.992	0.299	-1.163
10	-1.429	-0.777	-0.633	-1.388	-2.332	0.875	1.273	-1.722
11	-1.538	-0.339	-1.515	-0.315	-1.599	0.000	0.460	**-2.058**
12	-0.315	0.485	-0.649	-1.718	-1.414	0.700	**2.023**	-1.811
13	-0.622	-0.325	-1.906	0.000	-0.852	1.003	0.449	-0.211
14	-0.523	1.091	-0.086	0.070	**-2.317**	0.995	0.353	-0.686
15	-0.174	0.319	-0.090	-0.846	-1.886	1.395	1.144	-0.562
16	-1.365	-0.316	-1.273	-1.307	-1.324	0.636	0.508	**-2.135**

TABLE 2 T-tests

Danish data			
	asc1	*tc*	*tt*
Commuters			
2	**2.065**	0.955	0.132
3	**2.209**	1.096	-0.618
4	**2.214**	**2.402**	-1.013
5	**2.324**	1.227	0.506
6	**2.246**	**2.112**	-0.283
7	**2.419**	0.483	-0.545
8	**2.363**	0.547	0.371
Non-commuters			
2	**2.375**	-0.712	0.937
3	**2.605**	0.618	0.000
4	**2.731**	0.000	0.786
5	**3.067**	1.228	1.338
6	**3.060**	0.523	0.331
7	**2.910**	2.281	1.176
8	**3.298**	**1.966**	1.278

Fungibility data						
	cost	*safety*	*CV_first*	*cost_vs_safety*	*time_vs_safety*	*time*
2	0.242	0.000	0.789	-1.578	-0.412	-1.585
3	0.843	-0.715	0.761	-1.311	-0.872	-1.300
4	0.885	0.000	0.494	-1.046	-0.765	-1.049
5	0.653	-0.795	0.052	-1.216	-0.444	-1.220
6	0.377	0.000	1.488	-1.825	-1.594	-1.837
7	0.605	0.000	0.524	-1.470	-0.572	-1.465
8	1.602	-1.118	0.080	-1.849	-1.296	-1.840
9	0.742	-1.030	1.279	-1.731	-1.215	-1.736
10	1.190	-1.106	0.108	-1.389	-0.854	-1.393
11	1.237	-1.622	0.107	-1.687	-1.338	-1.689
12	**2.247**	**-2.366**	0.415	-1.437	-1.550	-1.430
13	1.370	-1.868	-0.053	-1.874	-1.469	-1.874
14	1.888	**-2.266**	-0.477	-1.604	-1.615	-1.603
15	1.854	**-2.218**	0.514	-1.831	-1.618	-1.842

TABLE 3 T-tests

Ecological value of Polish forests data										
	asc1	*asc2*	*asc3*	*nat1*	*nat2*	*tra1*	*tra2*	*inf1*	*inf2*	*fee*
2	**-2.602**	**23.192**	**-8.864**	**-6.175**	**-6.312**	**-3.337**	**-2.758**	**4.006**	**-2.238**	-1.855
3	0.100	**14.936**	**-10.569**	**-3.962**	**-2.456**	-1.548	-1.078	**3.297**	**3.044**	0.362
4	**-3.546**	**19.832**	0.000	**-5.624**	**-4.893**	**-4.542**	**-3.831**	**-3.148**	**-4.336**	**-6.185**
5	**-2.446**	**21.180**	**21.113**	**-6.541**	**-6.785**	**-4.452**	**-2.282**	0.892	**-2.493**	-0.184
6	**-2.436**	**10.190**	**-7.279**	**-6.704**	**-6.246**	**-3.757**	**-2.459**	**4.462**	-0.955	-1.547
7	**-3.545**	**23.111**	**21.043**	**-7.540**	**-7.353**	**-4.177**	**-4.057**	**2.560**	**-2.857**	**-2.364**
8	**-3.244**	**22.686**	**19.802**	**-7.131**	**-5.750**	**-4.878**	**-3.622**	-1.439	-1.743	**-4.068**
9	**-3.791**	**11.543**	**-3.906**	**-4.814**	**-4.966**	**-4.821**	**-4.154**	**-3.241**	**-4.460**	**-4.604**
10	**-2.436**	**23.428**	**9.818**	**-8.977**	**-5.291**	**-3.451**	**-2.735**	-0.276	0.217	-0.429
11	**-4.097**	**4.105**	**-10.298**	**-4.788**	**-7.099**	**-5.224**	**-4.715**	-1.426	**-2.435**	**-2.788**
12	**-2.905**	**9.328**	0.000	**-6.837**	**-6.755**	**-3.774**	**-1.993**	1.463	**-2.617**	-0.305
13	**-3.425**	**-2.141**	**-7.169**	**-5.028**	**-4.216**	**-4.277**	**-3.650**	**-3.663**	**-4.182**	**-7.459**
14	**-4.317**	**7.150**	**-8.180**	**-6.684**	**-5.398**	**-5.684**	**-4.092**	**-3.563**	**-4.502**	**-4.233**
15	**-3.966**	**2.130**	**-7.985**	**-5.695**	**-5.345**	**-5.026**	**-4.225**	**-4.216**	**-4.825**	**-5.229**
16	**-4.344**	**8.154**	**-8.551**	**-6.917**	**-5.910**	**-5.439**	**-4.587**	-0.989	**-3.798**	**-3.962**
17	**-4.000**	**13.603**	**-8.377**	**-9.072**	**-6.996**	**-5.277**	**-3.674**	0.834	**-3.229**	-0.731
18	**-3.975**	**19.240**	**-9.566**	**-7.481**	**-6.932**	**-4.409**	**-3.934**	0.000	**-3.236**	-1.780
19	**-4.412**	**11.124**	**-9.650**	**-9.025**	**-6.702**	**-5.472**	**-4.120**	**-2.620**	**-3.499**	-1.634
20	**-3.377**	**11.123**	0.000	**-8.611**	**-6.550**	**-4.483**	**-2.208**	**-2.966**	**-2.649**	**-2.029**
21	**-4.634**	**23.198**	**-4.579**	**-7.844**	**-6.957**	**-6.212**	**-5.154**	-0.892	**-4.166**	**-2.742**
22	**-4.461**	**16.913**	**-8.917**	**-7.897**	**-6.682**	**-5.123**	**-4.332**	**-3.338**	**-3.824**	**-2.997**
23	**-3.233**	**17.076**	**-7.732**	**-7.143**	**-6.326**	**-4.167**	**-3.520**	**3.401**	**-2.581**	-0.675
24	**-3.984**	**23.338**	**-3.439**	**-7.594**	**-5.908**	**-5.482**	**-4.002**	**2.123**	**-3.973**	**-2.487**
25	**-3.720**	**24.080**	**21.849**	**-7.887**	**-5.620**	**-5.610**	**-3.840**	**-3.179**	**-2.424**	-1.498
26	**-3.777**	**23.045**	**20.304**	**-7.754**	**-5.960**	**-5.784**	**-2.759**	**-2.428**	**-3.478**	**-2.504**

TABLE 4 T-tests

Białowieża Forest data							
	asc	*asc2*	*cen*	*gos*	*vis1*	*vis2*	*fee*
2	-0.911	-0.037	-0.964	-0.806	0.267	-0.789	-0.769
3	0.409	0.000	-0.344	0.627	1.060	-0.095	-0.778
4	0.405	-1.909	1.021	0.633	0.854	-0.363	-0.393
5	-0.565	-1.115	-0.106	-0.541	0.397	-1.193	-0.782
6	0.043	-1.096	0.356	0.273	1.900	-0.603	-0.380
7	-0.065	-0.442	-0.199	0.802	0.838	-0.425	-1.168
8	-0.824	-0.722	-0.750	-0.561	0.268	-0.235	-1.183
9	-0.461	-1.049	-0.195	-0.263	1.464	-0.660	-1.572
10	-0.043	0.112	0.174	-0.593	1.150	-0.534	-0.791
11	-1.030	-1.267	-0.475	-1.066	0.023	-0.704	-1.575
12	-0.727	-0.093	-0.051	-0.978	-0.040	-1.636	-1.175

Warsaw theatres dataset						
	asc	*roz*	*sro*	*dzi*	*eks*	*cost*
2	**-2.511**	**-2.445**	-1.424	-0.391	-0.784	-0.279
3	-0.476	-0.405	-0.701	-0.134	-1.229	0.700
4	-0.758	0.417	-1.707	0.000	-0.722	1.244
5	-2.050	-0.738	-0.628	-0.613	-1.517	-0.795
6	-0.413	-0.131	-0.409	0.542	-0.278	0.078
7	-1.865	0.267	**-2.181**	0.412	-1.252	0.660
8	-1.000	0.000	-1.666	-0.404	-0.137	0.777
9	-1.718	-1.352	-1.266	0.699	-1.405	1.097
10	-1.573	-1.451	-0.420	-0.399	-0.416	0.619
11	-0.533	0.286	-1.320	0.142	-0.715	1.836
12	**-2.261**	-0.402	-0.693	-0.673	-1.382	0.659

TABLE 5 Common Area of the kernel density estimates

AC statistic		Białowieża forest data							Second Australian dataset							
		sq	sq2	cen	gos	vis1	vis2	fee	asc_rp	asc_sp_2	cost	ct	ff	sdt	toll	var
Białowieża forest data	sq	1	0.40	0.44	0.51	0.50	0.63	0.29	.	.	.	.	.	.	.	.
	sq2	.	1	0.18	0.24	0.31	0.52	0.11	.	.	.	.	.	.	.	.
	cen	.	.	1	*0.69*	0.42	0.28	0.45	.	.	.	.	.	.	.	.
	gos	.	.	.	1	0.47	0.33	0.44	.	.	.	.	.	.	.	.
	vis1	.	.	.	.	1	0.33	0.37	.	.	.	.	.	.	.	.
	vis2	.	.	.	.	.	1	0.15	.	.	.	.	.	.	.	.
	fee	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.
Second Australian dataset	asc_rp	0.12	0.31	0.04	0.05	0.08	0.16	0.02	1	0.42	0.06	0.14	0.11	0.22	0.05	0.17
	asc_sp_2	0.37	0.70	0.16	0.21	0.29	0.46	0.10	.	1	0.23	0.43	0.36	0.56	0.22	0.48
	cost	0.48	0.25	0.47	0.52	0.64	0.29	0.47	.	.	1	0.42	0.63	0.24	0.47	0.50
	ct	*0.74*	0.49	0.40	0.46	0.45	0.66	0.24	.	.	.	1	0.58	0.56	0.53	0.61
	ff	*0.63*	0.39	0.49	0.56	0.70	0.46	0.33	.	.	.	.	1	0.40	0.53	0.66
	sdt	0.47	0.59	0.20	0.25	0.29	0.67	0.11	.	.	.	.	.	1	0.29	0.54
	toll	0.57	0.25	*0.67*	*0.71*	0.45	0.40	0.35	.	.	.	.	.	.	1	0.44
	var	0.62	0.50	0.36	0.43	0.58	0.55	0.26	.	.	.	.	.	.	.	1