Environmental and Resource Economics

Experimental Design Criteria and Their Behavioural Efficiency: An Evaluation in the Field

--Manuscript Draft--

Manuscript Number:	EARE-D-13-00119R2
Full Title:	Experimental Design Criteria and Their Behavioural Efficiency: An Evaluation in the Field
Article Type:	Manuscript
Keywords:	choice experiment; experimental design; latent class logit model; production forests; threatened native species
Corresponding Author:	Richard Tolentino Yao, PhD in Environmental Economics Scion (NZ Forest Research Institute) Rotorua, Bay of Plenty NEW ZEALAND
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Scion (NZ Forest Research Institute)
Corresponding Author's Secondary Institution:	
First Author:	Richard Tolentino Yao, PhD in Environmental Economics
First Author Secondary Information:	
Order of Authors:	Richard Tolentino Yao, PhD in Environmental Economics
	Riccardo Scarpa
	John Rose
	James Turner
Order of Authors Secondary Information:	
Abstract:	Comparative results from an evaluation of inferred attribute non-attendance are provided for experimental designs optimised for three commonly employed statistical criteria, namely: orthogonality, Bayesian D-efficiency and optimal orthogonality in the difference. Survey data are from a choice experiment used to value the conservation of threatened native species in New Zealand's production forests. In line with recent literature, we argue that attribute non-attendance can be taken as one of the important measures of behavioural efficiency. We focus on how this varies when alternative design criteria are used. Attribute non-attendance is inferred using an approach based on constrained latent classes. Given our proposed criterion to evaluate behavioural efficiency, our data indicate that the Bayesian D-efficiency criterion provides behaviourally more efficient choice tasks compared to the other two criteria.
Response to Reviewers:	Addressing the comments from the editor and reviewers (EARE-D-12-00119R1) Comments of editor and reviewers are in text without bullet points; responses of authors are reported in bullet points. We have also included an MS Word version of our responses in the set of attached files. EARE-D-12-00119R1 June 5, 2014 Dear Richard Yao, Thank you for submitting a revision of your paper, "Experimental Design Criteria and Their Behavioural Efficiency: An Evaluation in the Field" to Environmental & Resource Economics (ERE). I opted to send the paper out again for review, and now have heard back from the both of the original reviewers. The two reports are appended below.

I am pleased to say that both reviewers recommend acceptance of the paper subject to (a total of) three minor revisions.

•Thank you very much for your message and comments.

The one suggestion that warrants some thought is the request for some “discussion on weakness of using ANA as the measure of efficiency”. I ask that you address this.

•Thank you for pointing this out. A brief discussion on the weakness of ANA as the measure of efficiency is now written in Lines 65-74.

In reading your paper closely I have a few comments and suggestions that I would like you to incorporate. One major concern I have had with this study is the sample size. Please be explicit in the text that your analysis is based on three subsamples of 56 respondents (unless I misunderstood something). Of course, even if all respondents were under the same experimental design, it is often difficult getting a choice experiment published with less than 200 respondents. The sample size does open up the criticism of whether your results are subject to sampling error as it could simply be by chance that there are correlations between the design and the presence of ANA. I am not suggesting you need to go out and collect more data. But instead just appropriately caveat the findings. On a related point, one is usually concerned with the typical estimators for the variance-covariance matrix when the number of independent observations is small. Does your analysis account for this?

•We have now made it explicit in Lines 386-391 that we derived the 503 observations for each subsample from at least 56 respondents. We have now written that our total sample size was 172 respondents.

•To address your other concern, we have now written in Lines 394-397 that:

“The pooled sample size of 172 would appear small if no allowance is made for the high efficiency of the designs used in this application. However, we note here that the asymptotic properties of the estimator converge at the unusual rate of the square root of the sample size and should already be effective at this number of respondents.”

Here are some minor suggestions:
1.Abstract. Delete the word “contributions”.

•The word is now deleted.

2.Abstract. Perhaps state instead “optimal orthogonal in the difference design” to be clearer. When I read “orthogonal design” and “optimal orthogonal design” I wondered how these could possibly be different (i.e. orthogonal designs are of course based on optimality criteria).

•Thank you for this suggestion. We have now changed from “optimal orthogonality” to “optimal orthogonality in the difference” throughout the manuscript (e.g. Lines 7, 213). An orthogonal design is often not unique for a set of attributes and levels. The word “optimal” applies to the search for the most efficient of these orthogonal designs according to some a-priori and plausible assumption (e.g. the price coefficient should be negative, more is better, etc.)

3.Introduction. A snapshot of CE applications is a lackluster way to begin this paper. I would simply delete this and begin by motivating the research with discussion of the need for assessing the efficiency of competing experimental designs.

•Thank you for this suggestion. We have now deleted the snapshot and replaced it with the motivation of the research. Please see Lines 22 to 28.

4.Page 2. I am not sure what you mean by “theoretically valid framework”. It would be hard to argue that all your respondents are in fact revealing their true preferences. I suppose it is valid conditional on respondents actually making choices that maximize utility.

	•Thank you for this suggestion. We have now deleted those words as those might confuse the readers. 5.Page 3. Especially for the more casual reader, this discussion is not clear without at least a brief description of what you mean by serial ANA or the fully compensatory "assumption". •Thank you for pointing this out. We now explain both serial non-attendance and fully compensatory choice behaviour. Please see Lines 52-58. 6.Equation (6) should be reformatted as the lhs looks like D "minus" error. •Equation 6 now reformatted as suggested. Please see the row after Line 228. 7.The mathematical notation is not consistent throughout, e.g., the beta vector is only sometimes bolded. I recommend bolding vectors and matrices throughout. •Thank you for pointing out this oversight. All vectors and matrices are now in boldface font throughout. 8.First sentence of the conclusion: should be "design" rather than "designs". •Thank you for this suggestion. We have now changed "designs" to "design". 9.The discussion on pages 16-17 was a bit difficult to follow. If I understand correctly, you use the stated assessments of ANA to define possible latent classes (e.g. a cost ANA class), but you do not impose that a respondent that says they belong to a latent class to actually be in that class nor do you assign to them zero coefficients. Your approach makes sense, and avoids possible endogeneity concerns. But your discussion here can be condensed and what you do made more explicit. Perhaps place what others have done in a footnote. •You are correct, thank you for this suggestion. We have now rewritten Lines 331-345 accordingly and placed what others have done in Endnote number 4 (line 339), as suggested. 10.Page 17, middle paragraph. Delete "though,". •Thank you for this comment. "though" now deleted. At this point I am happy to recommend that your paper be accepted, conditional on addressing the remaining reviewer and editor comments. As I hope to simply accept your next revision "as is", I ask you to make sure that the paper adheres to the ERE style guidelines and that you go over the paper carefully to correct any remaining grammatical errors. •Thank you for this suggestion. We have gone through the paper thoroughly and carefully corrected the minor grammatical errors and to our eyes it now fully adheres to the ERE style guidelines. Thank you again for your submission. Best Regards, Christian Vossler Co-Editor, ERE Reviewer #1: Some minor issues: Update the reference Hole A (2011) A discrete choice model with endogenous attribute attendance. Economic Letters, 110(3), 203-205

	•Thank you for this suggestion. Reference now updated accordingly. Page 2, line 25 Louviere and Woodworth (2003). It is 1983, not 2003 •Thanks. “2003” now changed to “1983”. Reviewer #3: I appreciate the authors' responses and the improvement in clarity of the paper. I personally remain a bit skeptical of whether ANA is a "good" measure of behavioral efficiency (as opposed to a legitimate preference), but I agree with the author(s) that readers can make up their own mind and that some readers will agree and some will disagree. My only request is that you simply add some (small) discussion on weaknesses of using ANA as the measure of efficiency. •We have now elaborated on this (Lines 66-75) as requested. We have also added Endnote number 2 (Line 75) acknowledging and thanking an anonymous reviewer for this suggestion.

Manuscript resubmitted to

Environmental & Resource Economics

Experimental Design Criteria and Their Behavioural Efficiency: An Evaluation in the Field

Richard Yao*
Scion (NZ Forest Research Institute Ltd.)
Private Bag 3020, Rotorua, New Zealand
richard.yao@scionresearch.com
Phone: +64 7 343 5747; Fax: +64 7 348 0952

Riccardo Scarpa
The University of Waikato, Private Bag 3105
Hamilton, New Zealand
Gibson Institute, Queens University, Belfast, U.K.
CenSoC, University of Technology, Sydney, Australia
ric.scarpa@gmail.com

John Rose
University of South Australia Business School
Institute for Choice
New South Wales, Australia
John.Rose@unisa.edu.au

James Turner
AgResearch Ltd, Private Bag 3123
Hamilton, New Zealand
james.turner@agresearch.co.nz

*Corresponding author

Experimental Design Criteria and Their Behavioural Efficiency: An Evaluation in the Field

Abstract

Comparative results from an evaluation of inferred attribute non-attendance are provided for experimental designs optimised for three commonly employed statistical criteria, namely: orthogonality, Bayesian D-efficiency and optimal orthogonality in the difference. Survey data are from a choice experiment used to value the conservation of threatened native species in New Zealand's production forests. In line with recent literature, we argue that attribute non-attendance can be taken as one of the important measures of behavioural efficiency. We focus on how this varies when alternative design criteria are used. Attribute non-attendance is inferred using an approach based on constrained latent classes. Given our proposed criterion to evaluate behavioural efficiency, our data indicate that the Bayesian D-efficiency criterion provides behaviourally more efficient choice tasks compared to the other two criteria.

Key words: choice experiment, experimental design, latent class logit model, production forests, threatened native species

1 Introduction

The adoption of a given experimental design (ED) is often assumed to be behaviourally neutral in the practice of choice experiment (CE). However, the issue of whether technical features of the survey, such as the various types of ED, can be linked to systematic differences in structural parameter estimates has received very limited attention. This commonly held view is, therefore, no more than a plausible, yet uncorroborated assumption. In this paper we report on a study focussed on exploring this issue.

The arrangement of attribute levels for each alternative in a choice task is typically addressed by means of ED techniques. In a typical CE exercise, an analyst uses a single ED to derive the choice tasks presented to respondents as hypothetical scenarios consisting of a finite number of alternatives. Given standard assumptions, the minimum number of choice tasks required for the purposes of model estimation is a function of the number of attributes, attribute levels and alternatives in the choice tasks. Unfortunately, the number of attributes, levels and alternatives will often be such that the full factorial representing all possible combinations cannot be investigated in the survey. In such cases, analysts are faced with the challenge of selecting a fraction of the full factorial using some form of selection criteria. In order to elicit trade-offs, the alternatives in a choice task differ on a number of attribute dimensions and each respondent is typically expected to fully evaluate the utility consequences of these attribute level differences to select the preferred alternative in the choice task. This gives rise to the assumption of a fully compensatory choice behaviour that is in keeping with the random utility models used in estimation. The responses are then pooled to estimate utility weights of the sample of respondents for each of the design attributes (or attribute levels).

Most studies evaluating the performance of experimental designs for choice experiments investigate their statistical properties. The most commonly employed are various forms of statistical efficiency either using asymptotic (e.g., D-error, C-error, predictive efficiency, etc.) or finite sample approximations (Rose and Bliemer 2008, Yu et al. 2012). Explorations of other, arguably as important, behavioural components, such as some forms of 'behavioural' efficiency, are far less common. In this study we set out to investigate both statistical and behavioural performance of common ways of deriving experimental designs for stated choice surveys. Our analysis of the behavioural component focuses on inferred serial attribute non-attendance (IS-ANA), where serial non-attendance refers to the practice of some respondents to consistently ignore the same set of attributes when evaluating alternatives in a series of choice tasks. In the presence of systematic attribute non-attendance (ANA), the fully compensatory assumption commonly embedded in choice models (i.e. respondents trade-off all attributes levels in evaluating each alternative to execute the choice task) fails. Serial non-attendance is inefficient as it does not conform to conventional behavioural assumptions in choice; it hence introduces bias in estimation when it is ignored. ANA is derived from observed choice data and introduced in econometric models whose structure is informed by self-reported attribute non-attendance (SR-ANA). The self-reports are obtained from responses to debriefing questions collected in the survey.[1] The role of different experimental design criteria in determining ANA is explored by randomly assigning equivalent sub-samples of respondents to different ED treatments.

Some arguments can be made to critique the use of ANA as a measure of behavioural efficiency. This term is interpreted by us quite broadly and we note that our definition is based on adherence of actual behaviour to postulated assumptions. This is not dissimilar to the concept of robustness of results (in our case estimates) to crucial assumptions (in our case fully compensatory choice behaviour, which is undermined by ANA). It can be argued that other behavioural inefficiencies occur, for example, a constant error scale across respondents and choices in the sequence that can lead to other inefficient choice behaviour due to variation on the level of certainty in choice. We do not address them here, but we certainly suggest that the effect of ANA on these other forms of inefficient behaviour should also be investigated in the future.[2]

We compare and contrast three ED criteria. Firstly, we use one of the original ED criteria used for constructing CEs – the *orthogonality criterion* (Louviere and Woodworth 1983; Louviere and Hensher 1983). This has been the most widely used design criterion in linear multivariate models. It was first proposed for the statistical analysis of treatment effects in biological experiments, such as ANOVA studies. The orthogonality criterion generates fractional factorial designs that exhibit no correlation between each row of attributes levels and/or between columns of alternatives. (Orthogonal designs are described in detail in Louviere, et al. (2000) and Hensher et al. (2005a)). One advantage of this criterion is that the analyst does not need any *a priori* knowledge of the population parameter estimates. Therefore, the analyst can generate an orthogonal design by simply knowing the number of attributes, number of alternatives and number of choice tasks per respondent, without any assumption on the relative effects of attributes and levels on utility. However, while orthogonality is a desirable property for linear models, there is now ample evidence that selecting fractions of a full factorial by means of other criteria can outperform orthogonal designs in statistical terms when using logit specifications (Sandor and Wedel 2001, 2002, 2005; Kessels et al. 2006; Ferrini and Scarpa 2007; Scarpa and Rose 2008; Bliemer and Rose 2009; Vermeulen at al. 2011). These alternative criteria often require some plausible assumptions to be made on the relative magnitude and signs of the utility coefficients when these are expected to be different from zero, as well as on the specification of the final choice model. But the degree with which they outperform orthogonal designs in statistical terms is such that many analysts are ready to invoke the necessary assumptions (see for example Bliemer and Rose 2011), especially when only small samples are practicable. Orthogonal EDs are as efficient in logit models only when all coefficient attributes are equal to zero. However, if one indeed believes that utility coefficients are all equal to zero, this would imply equi-probability across alternatives, once the effect of the alternative specific constants is accounted for, and cause one to question why the investigation should take place at all. Despite a vast body of literature indicating the relative statistical inadequacy of orthogonal designs in stated choice survey data, the practice of their use is still deeply ingrained (e.g., Balcombe and Fraser 2011). For this reason we include this criterion in our investigation.

For a single design problem and a given fraction of the full factorial, there are often many possible orthogonal designs available. This suggests that given some assumptions on the range of values that are deemed to be likely for the utility coefficients, a search over the set of orthogonal designs can be performed to select the orthogonal fraction that displays the best statistical (and possibly behavioural) efficiency in that context. Furthermore, since only differences count in utility models, the orthogonalization should refer to the differences between attribute levels. Optimised orthogonal designs on the differences are thus orthogonal fractions of the factorial that have been selected with this concept in mind (see Street and Burgess 2007). This is the second design criterion used in our study.

One of the emerging criteria for selection from the full factorial is the Bayesian *D-error* minimization criterion (Sandor and Wedel 2001; Kessels et al. 2006, 2008; Ferrini and Scarpa 2007; Rose and Bliemer 2008; Bliemer and Rose 2010), which has been extended to increase in efficiency of estimates of welfare measures, such as marginal willingness-to-pay (WTP) (Scarpa and Rose 2008, Vermeulen et al. 2011). Note that this is different from the usual *D-efficiency* metric. Its statistical properties have been thoroughly investigated, but mainly by means of Monte Carlo simulations and other numerical or analytical techniques (Kessels et al. 2011; Bliemer and Rose 2009, 2010, 2013). This criterion has been attracting increased attention in non-market valuation of environmental goods in both one shot and multi-staged adaptive designs (Scarpa et al. 2007; Kerr and Sharp 2010), and we have chosen it as the third criterion subject of comparison in our empirical study.

Other studies investigate the behavioural efficiency of experiment design criteria in an empirical context, such as Bliemer and Rose (2011), Hess et al. (2008), Viney et al. (2005), Severin (2001) and Kinter et al. (2012). This type of efficiency may be just as important as statistical efficiency, since the quality of the model estimates depends on both forms. Overall, the joint gains in statistical and behavioural efficiency enable the analyst to reduce the required sample size and/or reduce the number of choice tasks necessary to achieve a given degree of precision in the estimation of the relevant parameters. This translates into a reduction in survey costs and in respondents completing surveys more quickly.

Whilst consensus on the measurement of statistical efficiency is well-established (Sandor and Wedel 2001, 2002, 2005; Scarpa et al. 2007; Ferrini and Scarpa 2007; Scarpa and Rose 2008), the measurement of behavioural efficiency is less well known, especially in systematic comparisons across designs. This makes it a more controversial issue. In this paper, we draw from a broad literature survey through which we identified a measure that has recently been attracting increasing attention. This is serial ANA, which is often interpreted as a behavioural response to the cost of cognitive effort and is predicated on the assumption that respondents are ‘cognitive misers’ (Fiske and Taylor 1984). As such, respondents would adopt decision heuristics that reduce their cognitive effort and tend to systematically switch off from considering the variation in levels of selected attributes (Campbell et al. 2008; Carlsson et al. 2010; Scarpa et al. 2009; Meyerhoff et al. 2009; Hensher and Greene 2010; Hole 2011; Scarpa et al. 2010; Balcombe et al. 2011; Hensher et al. 2012). Accounting for ANA has been found to have substantial effects on utility and welfare estimates in previous studies, with directions of bias that are not easy to predict *a priori*. Overall, it represents a form of inefficiency, the reduction of which is desirable. A desirable feature of a design criterion is the reduction of ANA effects. In this study we set out to empirically and systematically measure ANA effects across three experimental design criteria.

2 Design Efficiency in Choice Models

The Random Utility Maximization (RUM) framework proposed by Thurstone (1931), and later expanded upon by such researchers as McFadden (1974) and Manski (1977), provides the standard framework for modelling the choice behaviour of an individual. Under the RUM framework, an individual evaluates *J* alternatives in choice task *s* and selects the alternative that provides the highest expected utility. The usual utility function has deterministic and stochastic components as modelled by the basic conditional logit model. The analyst aims to estimate a $1 \times K$ row of utility weights or utility coefficients $\boldsymbol{\beta}$ for a column of vector $\boldsymbol{X}$ of $K \times 1$ attribute levels for respondent *n*'s indirect utility function. The utility function is usually expressed in a linear fashion as:

$$\boldsymbol{U}_{nj} = \boldsymbol{\beta}\boldsymbol{X}_{nj} + \boldsymbol{\varepsilon}_{nj} \qquad (1)$$

where $\boldsymbol{\varepsilon}_{nj}$ is the random error term that is i.i.d. Gumbel-distributed across *J* alternatives and *n* respondents. The conditional logit probabilities can be specified with the Gumbel error scale $\lambda > 0$ as:

$$\boldsymbol{P}_{nis} = \exp(\lambda(\boldsymbol{\beta}\boldsymbol{X}_{nis}))/\sum_{j=1}^{J} \exp\left(\lambda(\boldsymbol{\beta}\boldsymbol{X}_{njs})\right) \qquad (2)$$

where $\boldsymbol{P}_{nis}$ represents the probability that alternative i will be selected by respondent n from the set of J alternatives shown on choice task s. The values of X_{njs} are defined by the experimental design. A statistically efficient design is expected to maximise the amount of information the design conveys to identify the estimates for the vector of marginal utilities, $\boldsymbol{\beta}$. The information matrix for the design assuming a conditional logit model is defined by the matrix of second derivatives of the log-likelihood function presented as:

$$I(\boldsymbol{\beta}, \boldsymbol{X}_{njs}) = \frac{\partial^2 lnL}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} = \sum_{n=1}^{N}\sum_{j=1}^{J}\sum_{s=1}^{S} P_{njs}(\boldsymbol{X}_{njs} - \overline{\boldsymbol{X}}_{njs})(\boldsymbol{X}_{njs} - \overline{\boldsymbol{X}}_{njs})' \qquad (3)$$

where $\overline{\boldsymbol{X}}_{njs} \equiv \sum_{j=1}^{J} \boldsymbol{P}_{njs}\boldsymbol{X}_{njs}$

where $I(\boldsymbol{\beta}, \boldsymbol{X}_{njs})$ has a dimension of $K \times K$ which represents the Fisher Information Matrix (**FIM**). The **FIM** is a measure of the amount of information that observable sources of utility $\boldsymbol{X}_{njs}$ provide about $\boldsymbol{\beta}$ in explaining choice probabilities.

The conditional logit model takes a closed form (Train 2009) that implies a simple mathematical formulation of both the Jacobian (vector of first derivatives of the Log-likelihood function) and the Hessian (matrix of second derivatives of the Log-likelihood function). As these two matrices are functions of utility coefficients $\boldsymbol{\beta}$ and of the experimental design, $\boldsymbol{X}_{njs}$, an experimental design that increases the information embedded in the elements of **FIM** with respect to a baseline design is a more informative design. It is important to note that the negative of the inverse of the expected **FIM** is one of the maximum likelihood estimators of the asymptotic variance-covariance (**AVC**) matrix that can be shown as:

$$\boldsymbol{AVC} = \boldsymbol{\Omega}(\boldsymbol{\beta}, \boldsymbol{X}_{njs}) = -\left[E\left(\boldsymbol{I}(\boldsymbol{\beta}, \boldsymbol{X}_{njs})\right)\right]^{-1} = -\left[\frac{\partial^2 ln\boldsymbol{L}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right]^{-1} \tag{4}$$

where $ln\boldsymbol{L}$ is the log-likelihood of design $\boldsymbol{X}_{njs}$:

$$ln\boldsymbol{L} = \sum_{n=1}^{N}\sum_{j=1}^{J}\sum_{s=1}^{S}\boldsymbol{Y}_{njs} ln\boldsymbol{P}_{njs}(\boldsymbol{X}_{njs,}\boldsymbol{\beta}) \tag{5}$$

and $\boldsymbol{Y}_{njs}$ represents the indicator of choice that takes the value of 1 (if chosen) or 0 otherwise. The diagonal and off-diagonal elements of **AVC** represent, respectively, the variances and covariances of the elements of the $\boldsymbol{\beta}$ vector. The smaller the elements of **AVC** of the design, the more efficient the design is. A good criterion for choosing an efficient design is the one that minimises the determinant of the **AVC** matrix. An appropriate algorithm to generate and search for an efficient design would need to generate new designs from an initial coded design matrix, evaluate iteratively each new candidate design based on some criterion (e.g. efficiency) as a function of the arrangement of attribute levels, and identify the generated design that has an **AVC** with a sufficiently low determinant.

Scarpa and Rose (2008) described key measures of statistical efficiency of experimental designs used in modern choice experiments data collection that are often used to estimate non-linear models (e.g., logit). Two key types of experimental designs were described by Scarpa and Rose: one that assumes that all coefficients, $\boldsymbol{\beta}$, are equal to zero, and one that assumes otherwise. Street and Burgess (2004) developed an optimal experimental design under the assumption that the elements of $\boldsymbol{\beta}$ are all equal to zero. This is the assumption behind the optimal "orthogonal in the difference" criterion. We use the term D_z-*error* to represent the criterion's "efficiency" measure. However, in most practical cases the "$\boldsymbol{\beta}$ equal to zero" assumption might be considered too naïve. A choice analyst often spends a considerable amount of time identifying the attributes that are likely to influence utility, and would often have clear expectations as to the signs of their effects and hence of the coefficients. Additionally, in case of doubt, focus groups and conversations with experts in the field may be effective in identifying what would influence the utility experienced from the environmental good under study. Thus, one can expect that most, or even all, of the attributes would not equal zero. For example, at a minimum, in valuation experiments one could readily assume that the cost or price attribute would have a negative coefficient. This is informative as it rules out positive coefficient values.

The efficiency of the design that assumes (more realistically) that $\boldsymbol{\beta}$ values are not equal to zero is often measured by the *D-error*, which is based on the determinant of the **AVC** matrix of a design assuming a conditional logit model. This measure can be expressed as:

$$D\text{-}error = det\left(\boldsymbol{\Omega}\left(\boldsymbol{\beta}, \boldsymbol{X}_{njs}\right)\right)^{1/K} \qquad (6)$$

where in a choice experiment exercise, respondent *n* faces *J* alternatives, *K* attributes, and *S* choice tasks. As *K* increases, so does the number of elements in the $\boldsymbol{\beta}$ vector of indirect utility coefficients. This is accounted for by including the exponent $1/K$ in the equation. The term $\boldsymbol{\Omega}(\boldsymbol{\beta}, \boldsymbol{X}_{njs})$ represents the **AVC** matrix that is the negative inverse of **FIM**. This inverse relationship indicates that minimising the *D-error* leads to maximising the information of the experimental design. This suggests that the lower the *D-error*, the more informative, and hence statistically efficient the proposed design becomes, at least asymptotically.

Under the *D-error* set of assumptions, the values in $\boldsymbol{\beta}$ are treated with exact certainty. However, in reality, such values are uncertain. The Bayesian *D-error* (D_b) is an efficiency measure that accounts for uncertainty around the *a priori* values of $\boldsymbol{\beta}$. It can be expressed as:

$$D_b = \int \left[det\left(\boldsymbol{\Omega}(\boldsymbol{\beta}, \boldsymbol{X}_{njs})\right)\right]^{1/K} N(\boldsymbol{\mu}, \boldsymbol{\Sigma})d\boldsymbol{\beta} \qquad (7)$$

where the term $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ suggests that one may account for some *a priori* distributions of $\boldsymbol{\beta}$, which in our case is assumed to be normally distributed, with vector of means $\boldsymbol{\mu}$ and variance covariance $\boldsymbol{\Sigma}$. Ferrini and Scarpa (2007) suggested that less informative priors can also be invoked by assuming a uniform distribution. Under the D_b minimization criterion, it is typically assumed that utility coefficients are not equal to zero, but that uncertainty exists around the exact population values by assuming that such values are known only up to a distribution. Another scalar measure of design efficiency is the Bayesian *A-error* (A_b). In contrast to the determinant that accounts for all the elements of the **AVC** matrix, A_b only evaluates the trace, which is dependent only on the diagonal elements of the **AVC** matrix. As this measure does not account for the off-diagonals, this measure would likely provide higher scalar values than D_b. For this reason, D_b is more widely used than A_b in the experimental design literature.

3 Choice Behaviour Efficiency and Attribute Non-attendance

Attribute non-attendance is a processing strategy that can be employed by respondents in evaluating choice tasks. ANA is often thought to be the result of the simplifying heuristic strategies adopted by a respondent to reduce the cognitive cost of evaluating a series of experimentally designed choice tasks. Other processing strategies include: accounting for cost thresholds and cut-offs (Swait 2001; Han et al. 2001; Cantillo et al. 2006; Cantillo and Ortúzar 2006; Chou et al. 2008; Mørkbak et al. 2010; Campbell et al. 2012a); focussing on attribute levels previously experienced by respondents (Hensher 2008; Greene and Hensher 2010); and aggregating two different attributes (e.g. time and cost) into one on the basis of a common metric (Hensher 2006, Hensher and Layton 2008).

A number of CE studies have shown that some respondents, during the series of choice tasks they evaluate, tend to adopt choice behaviours involving ignorance of one or more attributes (e.g., Swait 2001; Hensher et al. 2005b, 2012; Hensher 2006, 2008, 2010; Fasolo et al. 2007; Islam et al. 2007; McIntosh and Ryan 2002; Lancsar and Louviere 2006; among others). In choice analysis, when ANA is suspected, it should be accounted for as its presence leads to the violation of the continuity axiom. This axiom implies that the choice model assumes fully compensatory choice behaviour from respondents, suggesting that they had attended to all attributes in a choice task (see Hensher 2006 for details of this axiom), otherwise changes in value levels of one attribute cannot be compensated with changes in value levels in another. In addition, accounting for different non-attending behaviours by respondents may contribute to significant improvements in goodness-of-fit measures[3] and more accurate or plausible estimates of welfare values (Scarpa et al. 2009, 2010).

Since we can now account for and detect the presence of ANA in choice data, we can also use it as a measure of behavioural efficiency of responses. We propose that a measure of ANA, such as the probability with which single attributes are predicted to be systematically ignored in the observed sequence of choice responses, is inversely related to behavioural efficiency. Sets of choice tasks with lower occurrence of ANA provide analysts with data that have been derived in a more considered manner and that are better aligned with standard application of choice models. This is because the more attributes attended to by respondents, the better the data satisfy the axiom of fully compensatory choice deliberation. Given that different experimental design criteria have different objectives (e.g., orthogonality restrictions, maximum D-efficiency, minimum *D-error*, etc.), in this study we explore whether or not choice tasks derived from different EDs criteria have varying levels of ANA. To do so, we analyse a balanced sample with split designs using the latent class logit approach to model inferred ANA (see also Scarpa et al. 2013). If ANA varies across designs, the design criteria that generated the series of choice tasks with the lowest occurrence of non-attendance to attributes would be considered as the most behaviourally efficient. It is worth mentioning that other studies have looked at other forms of inefficiency in choice behaviour. For example, Louviere et al. (2008) found that increased statistical efficiency as measured by D-efficiency (not *D-error* minimization) was correlated with a marked decrease in choice consistency (a form of behavioural efficiency) as measured by the relative size of the scale parameter of the Gumbel error.

4 Inferring ANA and Implementing It From Self-reports

Empirical evidence presented by Scarpa et al. (2009) showed different types of ANA behaviour where some respondents ignored one attribute, others ignored more than one and a few ignored all attributes (a choice behaviour consistent with random choices). Their results suggest that accounting for different types of non-attending behaviour of respondents contributes to a significant improvement in model goodness of fit and to more accurate estimates of parameter values. These authors suggested a modelling technique that allows the grouping of respondents (up to a probability) into different latent classes that could represent groupings based on non-attendance to certain subsets of attributes.

We can infer ANA from patterns of observed choices by using a panel Latent Class Logit Model (ANA-LCM) as described in Scarpa, et al. (2009). Conditional on belonging to a given ANA class, and therefore a given pattern of attended and not attended attributes, $\boldsymbol{\beta}_c$, the probability of observing the sequence of choices $\boldsymbol{Y}_n$ is defined as:

$$\boldsymbol{P}_n(\boldsymbol{Y}_n|\boldsymbol{\beta}_c) = \boldsymbol{P}_s(i_1, i_2, \dots, i_S|\boldsymbol{\beta}_c) = \prod_{s=1}^{S} \frac{\exp(\boldsymbol{X}_{is}\boldsymbol{\beta}_c)}{\sum_j \exp\left(\boldsymbol{X}_{js}\boldsymbol{\beta}_c\right)} \qquad (7)$$

where c represents latent classes formulated in terms of non-attendance, $\boldsymbol{P}_n$ represents the probability of respondent n observing a set of S choices, and $\boldsymbol{Y}_{n=\{y_1, y_2, \ldots, y_S,\}}$ is a product of logits $\prod_{s=1}^{S} \frac{\exp(\boldsymbol{X}_{is}\boldsymbol{\beta}_c)}{\sum_m \exp(\boldsymbol{X}_{js}\boldsymbol{\beta}_c)}$. To obtain the unconditional probability of the panel of choices of respondent n, the law of total probability is used. This is achieved by summing the conditional probabilities over the finite set of membership probabilities, $P(c)$, for each of the postulated ANA classes. The unconditional probability can be expressed as:

$$\boldsymbol{P}_n(\boldsymbol{Y}_n) = \sum_c \boldsymbol{P}(c)\, \boldsymbol{P}_n(\boldsymbol{i}_s|\boldsymbol{\beta}_c) = \sum_c \frac{\exp(\alpha_h)}{\sum_c \exp(\alpha_c)} \prod_{s=1}^{S} \frac{\exp(\boldsymbol{X}_{is}\boldsymbol{\beta}_c)}{\sum_j \exp\left(\boldsymbol{X}_{js}\boldsymbol{\beta}_c\right)} \qquad (8)$$

where α_h represents class-specific constants identified by some linear restriction (e.g., Latent Gold Choice imposes that they sum to zero (Vermunt and Magidson (2005)), whereas Nlogit imposes that one class has α_h=0 (Econometric Software, Inc. (2012)).

In the ANA-LCM above, the concept of ANA is operationalized by allowing individuals to be classified into latent behavioural classes. In each of these non-attendance classes some utility coefficients for attributes are restricted to zero, which is the value consistent with the utility effects of attributes that are not attended to, and hence not traded-off with others. The coefficients of those attributes that are attended to are, instead and obviously, allowed to be non-zero but are constrained to have exactly the same value across classes. In this sense, the classes differ across by indicating different attendance behaviour rather than taste heterogeneity, as is the case in conventional uses of latent class models. We assume that the specific structure of latent classes may be informed by self-reported statements of ANA. This is different from using a self-reported ANA statement on attributes in order to set the coefficients of the individual utility function to zero, as it is commonly done with self-reported ANA data; it also gets around, at least in part, the issue of endogeneity.[4] Previous studies on latent classes may also be used to identify which latent classes to include for testing. Suppose the identified and tested set of latent classes represents an adequate specification for our sample data, then the statistical fit of the model should significantly increase (relative to the conditional logit model) indicating not only the presence of non-attendance (suggesting that both a panel structure and discontinuous preference exists), but also that the non-attendance is well represented by using that latent structure. For comparisons of fit to the data, and to identify the most applicable number and types of latent classes (e.g., class ignoring the cost attribute, class ignoring the non-bird attributes composed of plant, lizard and fish), we use the minimum Akaike Information Criterion (AIC) approach (Swait 1994; Boxall and Adamowicz 2002). AIC is one of the alternative measures of goodness of fit to pseudo R^2 in non-linear regression models (e.g., conditional logit). Under the conditional logit model, AIC minimizes $-2\ln L + 2p$ where $\ln L$ represents the log-likelihood value and p is the number of parameters (Kennedy 2008). The smaller the AIC value the better the model fit while accounting for the number of parameters estimated. Estimation of the panel latent class logit models was undertaken using Latent Gold Choice software (Vermunt and Magidson 2005).

5 Data

The choice data were collected from a survey conducted between November 2009 and August 2010 (see Yao et al. (2014) for details). Three survey enumerators able to speak with New Zealand accents were employed to randomly telephone and invite more than 2,000 New Zealand individuals to participate in the phone-mail survey. Those who agreed in the phone screening to take part in the survey were sent a package containing the questionnaire, a return envelope, pen and pad. The sequential survey method of sending the surveys in two waves was used to improve operational conditions as described in Scarpa et al. (2007). The experimental design technique used for the first wave followed the orthogonal design (ORD) methodology. The ORD was composed of 27 choice situations divided into three blocks. Each respondent was given nine choice tasks to evaluate, each of which had three alternatives inclusive of the *status quo* (SQ) and two experimentally designed hypothetical and alternative states. The SQ alternative represented the current situation available at zero cost, while the other two represented changed forest states whose combination of levels were generated using the NGENE software (ChoiceMetrics 2012) for experimental design. Each alternative forest state was described by means of six attributes. The first five attributes consisted of three levels of occurrence or abundance of threatened species in New Zealand planted forests (Table 1). The sixth attribute was the cost defined in four levels of additional annual income tax for five years ($0, $30, $60 and $90). The attributes and their respective levels and dummy coding used in estimation are shown in Table 1; an example of a choice task used in the survey is presented in Figure 1.

[Table 1 goes about here]

[Figure 1 goes about here]

In the second wave of the survey, as well as ORD, two more EDs were included: a Bayesian D-efficient (BDD) and an optimal orthogonal in the difference design (OOD) (Street and Burgess 2004, Street et al. 2005). In generating the BDD and OOD, we assumed that the choice data collected would be analyzed using a conditional logit model. As in the first wave, BDD and OOD were generated using NGENE. To generate BDD choice tasks, we used the conditional logit model estimates from the first wave of survey completed by 35 respondents, to derive the *a priori* distribution of the parameters of the indirect utility function (Appendix Table 1). To generate the designed alternatives for OOD, an *a priori* assumption is unnecessary.

From the first and second waves of survey, we derived a balanced sample of 1,509 choice observations that were evenly distributed across the three EDs. For an objective comparison of the three design treatments, we allocated 503 choice observations derived from at least 56 respondents per treatment to each design sample (Table 2). The pooled sample size of 172 would appear small if no allowance is made for the high efficiency of the designs used in this application. However, we note here that the asymptotic properties of the estimator converge at the unusual rate of the square root of the sample size and should already be effective at this number of respondents. All three choice sub-samples have equal numbers of observed choice task orders (i.e., 56 observations for the 1^{st}, 2^{nd}, 4^{th}, 5^{th}, 6^{th}, 7^{th}, 8^{th} and 9^{th} choice task orders; and 55 observations for the 3^{rd} choice task order) (Table 2). To construct a balanced sample and complete allocation to treatments, we have excluded a few choice observations in the OOD and ORD samples to facilitate consistency with the BDD sample.[5] We excluded 9 choice observations using the following criteria: (1) if respondents did not complete the nine choice tasks; (2) if respondents sent back the questionnaire too late; and (3) for convenience, other choice observations at the bottom of the worksheet were removed when in excess of the balance required by the design. A sensitivity analysis showed that the deletion of those specific choice observations, rather than others, to balance the treatments, did not change the salient results.

[Table 2 goes about here]

The ORD sample includes all choice observations from the first wave (35 respondents), with the rest the second wave. Choice data for the BDD and OOD samples were collected from the second wave of survey only.

6 Evaluation of the Experimental Designs

Each choice task was checked for the presence of dominant alternatives before using the BDD as designed by the software NGENE to collect the survey data. With the assumption that the utility of an individual increases monotonically with the improvement in attribute levels (i.e., Level 2 is strictly preferred to Level 1 which is strictly preferred to the current condition), two choice tasks were found with dominant alternatives in one of the three blocks. As conventionally done in practice, we eliminated the presence of dominance in the BDD by swapping attribute levels across choice tasks within a block. Although this procedure minimally affected the design efficiency, it was felt necessary to eliminate dominant choice tasks as suggested in Greene and Hensher (2003) (see also Kessels et al. 2011 for a discussion of the implications of retaining such choice tasks) and to emulate the state of practice in the field. The results of the evaluation of the statistical efficiency of the three final designs following the design efficiency measures in Scarpa and Rose (2008) and Street and Burgess (2004) are given in Table 3. As can be expected, OOD has the lowest D_z-error and A_z-error implying that OOD is the most efficient design under this measure. For the second set of measures, where we assumed that parameter values were to be based on *a priori* information (i.e., $\boldsymbol{\beta}_S \neq \mathbf{0}$), the BDD is the most efficient design based on the D_p and A_p criteria, while OOD has the lowest efficiency. This is unsurprising, as the BDD criterion produces the design that maximizes the value of the elements of the information matrix calculated on the basis of the coefficient estimates from the pilot data (from first wave of survey). These sets of priors can be considered valid because they came from actual survey respondents. Nevertheless, in view of the conclusions reported in Ferrini and Scarpa (2007), we elected to test whether the pilot data provided reliable priors once the full data became available. We employed the method described in Scarpa et al. (2005) where we compare estimated marginal WTPs between the pilot sample (WTP_P) and the full sample (WTP_F). Percentage differences in WTPs between attributes for the two sample groups are provided in Table 4. Level 2 (denoting an increase in abundance of *Brown kiwi*) is approximately nine percent lower in the full sample compared with the pilot sample, while the Level 1 increase in *Bush falcon* abundance is lower by about 28 percent. These relatively small WTP differences in key attributes between the pilot and full samples (provided Gumbel scale was the same across) suggest that our set of priors may be considered reliable. The WTPs for most non-bird attributes were not compared because of the statistically insignificant utility coefficients from the pilot sample.

[Tables 3 and 4 go about here]

7 Results

The estimates of conditional logit models for the three subsamples subject to the three design treatments are reported in Table 5. Cost coefficient estimates are all negative and significant, as expected. All statistically significant coefficients for the environmental attributes (e.g., *Brown kiwi 1*, *Brown kiwi 2*, *Bush falcon 2*) have positive signs, implying that increasing the abundance of these threatened species contributes positively to the utility of an individual. Some coefficient estimates (e.g., *Green gecko 1*, *Kakabeak 1*) have unexpected negative signs, but these are not statistically significant. Coefficient estimates for all non-bird species in the OOD sample are not statistically significant. These are species considered to be less charismatic and iconic than the Brown kiwi and the Bush falcon. As such we conjecture that they are more prone to suffer from non-attendance in our sample. Note that in this specification, the pseudo R^2 values show best fit for the model estimates on the ORD design, followed by the OOD and with the BBD displaying worst fit.[6] The BBD and ORD designs produce the largest number of attribute coefficient estimates significant at conventional values (ignoring the SQ), with the BBD data displaying most information in the Fisher information matrix at convergence. This confirms the highest efficiency of this design criterion in practice.

[Table 5 goes about here]

A summary of the proportion of respondents who self-reported ignoring at least one of the attributes while evaluating the choice tasks is presented in Table 6a. This question was asked after each respondent completed all nine choice tasks. These are the self-reported serial attribute non-attendance (SR-ANA) scores. The pooled sample shows a pattern that is consistent with at least one-out-of-ten respondents having ignored one non-bird attribute. As expected, the more iconic bird attributes had much lower non-attendance, with the highest frequency observed in the ORD sub-sample. The lowest SR-ANA score for non-bird attributes is shown for the sub-sample from the BDD criterion. Based on this SR-ANA information, we identified non-attendance to non-bird attributes as a candidate latent class for evaluating the behavioural performance of our design criteria by means of IS-ANA.

[Table 6a goes about here]

We note that the alternative specific constant (ASC) for SQ under the ORD design criterion is positive and significant, but not so for the other two designs (Table 5). We conjecture that respondents with choice tasks generated with the ORD criterion were more likely to choose the SQ alternative, implying that they have a higher tendency to opt out compared to respondents facing the other two designs. We would like to point out here that "opting out" can also be considered as a legitimate "real life preference" rather than a "bias", when all other alternatives are not sufficiently attractive. We investigate the SQ bias conjecture by including a second behavioural latent class in our IS-ANA model in which the SQ coefficient is restricted to zero.

The third behavioural latent class in our IS-ANA model is derived from Campbell et al. (2008) where it is suggested that 70 percent of the respondents might have ignored the cost attribute. Although the results in that paper might represent an extreme case, attendance to cost is important because in hypothetical valuations there is no penalty to respondents for ignoring price. On the other hand, accurate estimates of the price coefficient are important to researchers to obtain valid welfare estimates.

The fourth candidate latent class for our IS-ANA model is based on the conventional assumption that respondents attended to all attributes in evaluating choice tasks, hence behaving in a fully compensatory fashion. This full attendance class should be dominant in our data based on our SR-ANA scores where majority of respondents appear to have attended to all five environmental attributes (Table 6b). We also found that 70 percent of respondents stated they attended to all species used to describe the forest management scenarios (Table 6b). The design derived from the BDD criterion has the highest proportion of respondents self-reporting a fully compensatory choice (73 percent), closely followed by the OOD and ORD.

[Table 6b goes about here]

The estimates of the ANA-LCM for the three designs are provided in Table 7. This model is the tool from which we derive the IS-ANA model. To objectively compare the three design treatments, different combinations of the four candidate latent classes mentioned above were tested. These are: (1) full attendance; (2) ignored non-bird species; (3) ignored SQ; and (4) ignored cost.[7] As expected, the goodness of fit measures for all design treatments substantially improved from those in the conditional logit model when the latent class panel model is fitted to the choice data. For example, the log likelihood values for the ORD went from -459 to -265 with only four more parameters, with similar improvements for the other two designs. This provides strong evidence of the presence of heterogeneity in the specific form of attribute non-attendance across the three design treatments and the panel data nature of the observed choices.

[Table 7 goes about here]

Our results show that for the three ED treatments, respondents who evaluated choice tasks from the BDD have the highest probability (0.236) of belonging to the class with full attendance compared to the OOD (0.219) and ORD (0.010) (Table 7). This indicates, based on our data, that the BDD gave rise—everything else being equal—to a greater proportion of respondents attending to all attributes and thus producing choices consistent with the conventional assumption of fully compensatory behaviour. Importantly, this lower inferred non-attendance is consistent with the lower self-reported scores summarised in Tables 6a and 6b that show that relatively smaller proportion of respondents ignored choice attributes when faced with choice sets from the BDD design, compared to the two other designs. We are reluctant to provide an explanation for such a comparatively different result in both stated and inferred ANA in the BDD design as it would be exclusively speculative in nature at this stage. If it had been found only in the inferred ANA case, one could argue that it could be a property of the geometry of the design matrix. However, the fact that it was also associated with lowest stated ANA warrants further attention. This topic should be the focus of further research.

The ORD had the lowest membership probability to the latent class with full attendance, which reinforces the importance of using optimised experimental designs in choice modelling. We find that, with reference to between design treatments, the ORD displays the highest membership probability (0.297) to the class that ignores the non-bird attributes, while BDD and OOD assign a significantly lower membership probability (0.108 and 0.097, respectively) to this class. This may indicate that a larger proportion of respondents to these two designs had carefully accounted for both iconic and non-iconic species before selecting the preferred alternative. The ORD treatment also has the highest membership probability to the class ignoring SQ (0.454), not so closely followed by the BDD (0.364) and OOD (0.319), respectively. On the plus side, and importantly for the derivation of welfare measures, the ORD has the lowest membership probability value (0.239) for the latent class that ignored the cost attribute followed by the BDD (0.292) and OOD (0.365). In terms of overall goodness of fit of the model to the data for the four latent classes, the OOD treatment exhibits the best overall fit with an adjusted pseudo R^2 of 0.672. When inferred ANA is allowed for, the number of insignificant coefficient estimates at the 10 percent level is reduced to three in ORD and four in BDD, while for OOD it is still high with six insignificant estimates. Finally, with regards to opting out, the ratio of estimates between SQ cost coefficient for BDD is more than twice the ratio in the OOD and more than 70 percent larger than in the ORD model, which suggests that a typical respondent who evaluated a BDD choice would be much less likely to opt out relative to ORD and OOD.

8 Conclusions

In this work, we have explored the performance of alternative design criteria for choice experiments in terms of one form of behavioural efficiency within a survey format. In line with recent literature, we argue that serial attribute non-attendance can be taken as an important measure of behavioural efficiency, and we have focussed on how it may systematically vary when alternative design criteria are used. Based on the sample of data examined here, we found some empirical evidence of the superiority of the Bayesian D-efficient design (BDD) relative to the orthogonal design (ORD) and to the optimal orthogonal in the difference design (OOD). In line with other studies, we have confirmed that a BDD is statistically more efficient, and add to the literature by finding that it is also behaviourally more efficient than the two other designs. This is indicated by a smaller Bayesian *D-error* and a greater proportion of respondents who are likely to attend to all attributes in the choice tasks, as well as less inclined to opt-out by choosing the SQ. Therefore, we conclude that among the three common criteria used in the derivation of experimental designs for stated choice, BDD provides choice tasks that induce respondent behaviour most consistent with the common assumption of fully compensatory choice. Importantly, for the practice of welfare estimate derivation from stated choice data, we find that the probability of inferred non-attendance to the cost attribute ranges between one-fourth in the ORD sample and one-third in the OOD sample, while BDD was in-between with 30 percent. Clearly, this set of results may be specific to our sample data. It is thus suggested that future studies evaluating different EDs should investigate if more efficient designs also induce a lower rate of attribute non-attendance systematically to enable this to be taken as an empirical regularity. Our results add evidence to the issue of non-neutrality of the choice of experimental design in stated choice data, in the sense that estimates seem to be affected by the choice of criteria used to derive the experimental design used in allocating attributes and attribute levels across alternatives within choice tasks.

The length of time it took a respondent to evaluate the sequence of choice tasks and make each single choice was not recorded in this study, in contrast to the work described in Rose and Black (2006) as well as in Campbell et al. (2012b). Choice task completion time and other behavioural clues on the information capture of alternative descriptors, such as eye-tracking may help explore other behavioural efficiency measures. We suggest that future studies on attribute non-attendance behaviour should also include an evaluation of the effect of time taken by respondents to choose in each choice task and of the eye-track patterns of respondents during choice execution. Several online survey packages (e.g., www.qualtrics.com) allow the recording of the number of seconds and/or minutes it took a respondent to browse through certain pages of the online questionnaire. Eye-tracking, by contrast, is likely to involve more expensive equipment as well as costly and specific interview settings, but might produce more valid measure of behavioural efficiency, especially if integrated with data on brain activity during choice (Weber et al. 2007), the use of which is even more expensive. Finally and crucially, in a methodology that finds its main motivation in the derivation of estimates of non-market values, future research should focus on the sensitivity of welfare estimates to alternative criteria for deriving experimental designs from their full factorial.

References

Balcombe K, Burton M, Rigby D (2011) Skew and attribute non-attendance within the bayesian mixed logit model. J Environ Econ Manag 62(3): 446-461

Balcombe K, Fraser I (2011) A general treatment of ‘don't know’ responses from choice experiments. Eur Rev Agric Econ 38(2): 171-191

Bliemer MCJ, Rose JM (2013) Sample size requirements for stated choice experiments. Transportation (February), 1-21

Bliemer MCJ, Rose JM (2011) Experimental design influences on stated choice outputs: an empirical study in air travel choice. Transp Res A - Pol 45:63-79

Bliemer MCJ, Rose JM (2010) Construction of experimental designs for mixed logit models allowing for correlation across choice observations. Transp Res B - Meth 46(3): 720-734

Bliemer MCJ, Rose JM (2009) Designing stated choice experiments: the state of the art. In: Kitamura R, Yoshi T, Yamamoto T (eds) The expanding sphere of travel behavior research. Selected papers from the 11th International Conference on Travel Behavior Research, Kyoto, 16–20 August 2009

Boxall PC, Adamowicz WL (2002) Understanding heterogeneous preferences in random utility models: a latent class approach. Environ Resour Econ 23(4): 421–446

Campbell D, Hensher DA, Scarpa R (2012a) Cost thresholds, cut-offs and sensitivities in stated choice analysis: identification and implications. Resour Energy Econ 41:401-417

Campbell D, Mørkbak MR, Olsen SB (2012a) Response latency in stated choice experiments: impact on preference, variance and processing heterogeneity. Paper presented at the European Association of Environmental and Resource Economists 19th Annual Conference in Prague, Czech Republic, 27-30 June 2012

Campbell D, Hutchinson WG, Scarpa R (2008) Incorporating discontinuous preferences into the analysis of discrete choice experiments. Environ Resour Econ 34(3):396-411

Cantillo V, Heydecker B, Ortúzar JD (2006) A discrete choice model incorporating thresholds for perception in attribute values. Transp Res B - Meth 40:807-825

Cantillo V, Ortúzar JD (2006) Implications of thresholds in discrete choice modelling. Transport Rev 26:667-691

Carlsson F, Kataria M, Lampi E (2010) Dealing with ignored attributes in choice experiments on valuation of sweden's environmental quality objectives. Environ Resour Econ 47:65-89

ChoiceMetrics (2012) Ngene 1.1.1 user manual and reference guide. http://www.choice-metrics.com

Chou HY, Lu JL, Fu C (2008) The study of price accept threshold for the “blue highway” tour of the North-East Region in Taiwan. J Marine Sci Technol 16:255-264

Econometric Software, Inc (2012) NLOGIT 5. Plainview, New York

Fasolo B, McClelland GH, Todd PM (2007) Escaping the tyranny of choice: when fewer attributes make choice easier. Market Theory Decis 7:13-26

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

Fiske ST, Taylor SE (1984) Social cognition. Addison-Wesley, Massachusetts

Greene WH, Hensher DA (2003) A latent class model for discrete choice analysis: contrasts with mixed logit. Transp Res B - Meth 37: 681-698

Greene WH, Hensher DA (2010) Ordered choices and heterogeneity in attribute processing. J Transp Econ Policy 44: 331-364.

Han S, Gupta S, Lehmann DR (2001) Consumer price sensitivity and price thresholds. J Retailing 77:435-456

Hensher DA (2006) How do respondents process stated choice experiments? attribute consideration under varying information load. J Appl Econom 21:861-878

Hensher DA (2008) Joint estimation of process and outcome in choice experiments and implications for willingness to pay. J Transp Econ Policy 42:297-322

Hensher DA (2010) Hypothetical bias, stated choice studies and willingness to pay. Transp Res B - Meth 44:735-752

Hensher DA, Greene WH (2010) Non-attendance and dual processing of common-metric attributes in choice analysis: a latent class specification. Empirical Econ 39:413-426

Hensher DA, Layton D (2008) Attribute referencing, cognitive rationalisation and implications for willingness to pay. Working paper, Institute of Transport and Logistics Studies, The University of Sydney.

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

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

Hensher DA, Rose JM, Greene WH (2012) Inferring attribute non-attendance from stated choice data: implications for willingness to pay estimates and a warning for stated choice experiment design. Transportation 39:235-245

Hess S, Smith C, Falzarano S, Stubits J (2008) Measuring the effects of different experimental designs and survey administration methods using an Atlanta managed lanes stated preference survey. Transport Res Rec 2049:144-152

Hole A (2011) A discrete choice model with endogenous attribute attendance. Economic Letters, 110(3): 203-205.

Islam T, Louviere JJ, Burke PF (2007) Modeling the effects of including/excluding attributes in choice experiments on systematic and random components. Int J Market Res 24:289-300

Kennedy P (2008) A guide to econometrics (6th edition). Wiley-Blackwell

Kerr GN, Sharp BMH (2010) Choice experiment adaptive design benefits: a case study. Aust J Agric Resour Econ 54:407-420

Kessels R, Goos P, Vandebroek M (2006) A comparison of criteria to design efficient choice experiments. J Marketing Res 43:409-419

Kessels R, Goos P, Vandebroek M (2008) Optimal designs for conjoint experiments. Comput Stat Data An 52(5):2369-2387

Kessels R, Jones B, Goos P, Vandebroek M (2011) The usefulness of Bayesian optimal designs for discrete choice experiments. Appl Stoch Model Bus 27(3): 173-188

Kinter ET, Prior TJ, Carswell CI, Bridges JFP (2012) A comparison of two experimental design approaches in applying conjoint analysis in patient-centered outcomes research: a randomized trial. Patient 5(4):279-294

Lancsar E, Louviere JJ (2006) Deleting 'irrational' responses from discrete choice experiments: a case of investigating or imposing preferences? Health Econ 15:797-811

Louviere JJ, Woodworth G (1983) Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data. J Marketing Res 20:350-367

Louviere JJ, Hensher DA (1983) Using discrete choice models with experimental design data to forecast consumer demand for a unique cultural event. J Consum Res 10:348-361

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

Louviere JJ, Islam T, Wasi N, Street D, Burgess L (2008) Designing discrete choice experiments: do optimal designs come at a price? J Consum Res 35:360-375

Manski C (1977) The structure of random utility models. Theor Decis 8: 229-254

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

McIntosh E, Ryan M (2002) Using discrete choice experiments to derive welfare estimates for the provision of elective surgery: implications of discontinuous preferences. J Econ Psychol 23:367-382

Meyerhoff J, Liebe U (2009) Discontinuous preferences in choice experiments: evidence at the choice task level. Paper presented at the 17th EAERE conference, Amsterdam, 24-27 June 2009

Mørkbak MR, Christensen T, Gyrd-Hansen D (2010) Choke price bias in choice experiments. Environ Resour Econ 45:537-551

Puckett SM, DA Hensher (2009) Revealing the extent of process heterogeneity in choice analysis: an empirical assessment. Transport Res A - Pol 431:117-126

Rose JM, Black I (2006) means matter, but variance matter too: decomposing response latency influences on variance heterogeneity in stated preference experiments. Marketing Letters 17(4):295-310

Rose JM, Bliemer MCJ (2008) Stated preference experimental design strategies. In: Hensher DA, Button KJ (eds) Handbook of Transport Modelling. Elsevier, Oxford

Sandor Z, Wedel M (2001) Designing conjoint choice experiments using managers' prior beliefs. J Marketing Res 38:430-444

Sandor Z, Wedel M (2002) Profile construction in experimental choice designs for mixed logit models. Market Sci 21:455-475

Sandor Z, Wedel M (2005) Heterogeneous conjoint choice designs. J Marketing Res 42:210-218

Scarpa R, Rose J (2008) Design efficiency for non-market valuation with choice modelling: how to measure it, what to report and why. Aust J Agric Resour Econ 52: 253-282

Scarpa R, Campbell D, Hutchinson GW (2007) Benefit estimates for landscape improvements: sequential Bayesian design and respondents' rationality in a choice experiment. Land Econ 83:617-34

Scarpa R. Zanoli R, Bruschi V, Naspetti S (2013) Inferred and stated attribute non-attendance in food choice experiments. Am J Agric Econ 95(1):165-180

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

Scarpa R, Gilbride TJ, Campbell D, Hensher DA (2009) Modelling attribute non-attendance in choice experiments for rural landscape valuation. Eur J Agric Econ 36:151-74

Scarpa R. Thiene M, Hensher DA (2010) Monitoring choice task attribute attendance in non-market valuation of multiple park management services: does it matter? Land Econ 86(4):817-839

Severin V (2001) Comparing statistical and respondent efficiency in choice experiments. PhD Dissertation, Discipline of Marketing, Faculty of Economics and Business, University of Sydney

Street DJ, Burgess L (2004) Optimal and near-optimal pairs for the estimation of effects in 2-level choice experiments. J Stat Plan Infer 118:185-199

Street DJ, Burgess L, Louviere JJ (2005) Quick and easy choice tasks: constructing optimal and nearly optimal stated choice experiments. Int J Res Mark 22:459-470

Street DJ, Burgess L (2007) The construction of optimal stated choice experiments: theory and methods. Wiley, New Jersey

Swait J (2001) A non-compensatory choice model incorporating attribute cutoffs. Transp Res B - Meth 35:903-928

Swait J (1994) A structural equation model of latent segmentation and product choice for cross-sectional revealed reference choice data. J Retailing Consumer Services 1:77-89

Thurstone L (1931) The indifference function. J Soc Psychol 2:139-167

Train KE (2009) Discrete choice methods with simulation. Second edition. Cambridge University Press

Vermeulen B, Goos P, Scarpa R, Vandebroek M (2011) Bayesian conjoint choice designs for measuring willingness to pay. Environ Resour Econ 48:129-149

Vermunt JK, Magidson J (2005) Technical guide for Latent GOLD Choice 4.0: basic and advanced. Statistical Innovations Inc., Massachusetts

Viney R, Savage E, Louviere JJ (2005) Empirical investigation of experimental design properties of discrete choice experiments in health care. Health Econ 14:349-362

Weber B, Aholt A, Neuhaus C, Trautner P, Elger CE, Teichert T (2007) Neural evidence for reference-dependence in real-market-transactions. NeurImage 35:441-447

Yao RT, Scarpa R, Turner JA, Barnard TD, Rose JM, Palma JHN, Harrison DR (2014) Valuing biodiversity enhancement in New Zealand's planted forests: Socioeconomic and spatial determinants of willingness-to-pay. Ecol Econ 98: 90-101

Yu J, Goos P, Vandebroek M (2012) A comparison of different Bayesian design criteria for setting up stated preference studies. Transp Res B - Meth 46:789-807

Endnotes

[1] As this study focuses on “serial ANA”, we asked each respondent the attribute or attributes that she/he ignored after evaluating all the choice tasks. Other CE studies also examined “choice task specific ANA” where each respondent was asked for the ignored attribute/s after evaluating each choice task (e.g. Hensher, 2006; Puckett and Hensher, 2009; Scarpa et al. 2010).

[2] We are thankful to an anonymous reviewer for suggesting to elaborate on other forms of behavioural inefficiencies worth investigating.

[3] It is also possible that accounting for ANA may result in poorer model fits. If, for example, a respondent is observed to always select the highest priced alternative over repeated choice tasks, under maximum likelihood estimation techniques, the model will naïvely assume that the respondent prefers higher priced products, thus assigning a positive parameter to that individual. If, in accounting for ANA, the respondent is assigned a parameter of zero (under the assumption that they ignored price), then a poorer model fit is likely to be observed. Mathematically, a better model log-likelihood will be obtained if the parameter were allowed to be positive as opposed to being constrained to be zero as a positive parameter will better match the observed data. As such, care is required when selecting specifications based only on model fit criteria.

[4] We note that self-reported statements of ANA can be directly implemented in choice models in a much simpler way, although we do not do it here. If respondent n self-reported ANA for attribute k, then this attribute will have a β_{kn} coefficient restricted to zero. This implementation is discussed in Hensher, Rose and Greene (2005a) and in

Campbell et al. (2008), amongst others. Similar to many previous studies that employed self-reported ANA, for its identification during the survey, we used a single de-briefing question posed to the respondent after the evaluation of all choice situations.

[5] Note that, even though we have excluded observations here to help facilitate the statistical tests to be performed, we do not recommend doing this in practice, particularly when using orthogonal designs. Orthogonality requires that each task in the design is equally replicated in a data set. Removing observations will induce correlations and hence destroy the properties of the design.

[6] Care should be taken, however, in putting excessive reliance on such comparisons because the log-likelihood function is data-specific. The concept of model fit provides little information in this context, as the data, and hence models, are non-nested.

[7] While we have also estimated specifications with classes, (e.g., ignoring the cost attribute, ignoring all attributes, and attending only to one attribute) our analysis indicates that this set of latent classes is the most suited to our pooled data set as it results to the lowest normalised AIC (AIC/n) value from among 10 other model specifications we employed in the grid search exercise (see Appendix Table 2).

Threatened Animal/Plant		*Current Condition*	*Option A*	*Option B*
Brown Kiwi (Frequency of hearing calls in planted forests in North Island)		Kiwi calls heard **in 1 out of 200** planted forests	Kiwi calls heard **in 20 out of 200** planted forests	Kiwi calls heard **in 1 out of 200** planted forests
Giant Kokopu (Occurrence in slow moving streams with overhanging native vegetation in planted forests throughout New Zealand)		Kokopu seen **in 1 out of 10** suitable streams	Kokopu seen **in 3 out of 10** suitable streams	Kokopu seen **in 1 out of 10** suitable streams
Kakabeak (Occurrence in 20% of the planted forests on the East Coast and Hawke's Bay)		At least **3 naturally occurring** Kakabeak shrubs	At least **3 naturally occurring** Kakabeak shrubs	At least **10 actively managed** Kakabeak shrubs
Auckland Green Gecko (Gecko sightings in open grounds in planted forests in Northland, Waikato and Bay of Plenty regions)		Gecko sighted **in 1 out of 50** walks	Gecko sighted **in 5 out of 50** walks	Gecko sighted **in 1 out of 50** walks
NZ Bush Falcon (Bush falcon sightings while driving through pine forests in Central North Island and Nelson)		Bush falcon sighted **in 1 out of 8** drives	Bush falcon sighted **in 3 out of 8** drives	Bush falcon sighted **in 1 out of 8** drives
Additional amount to be paid yearly in your income tax for five years only		**$0**	**$30**	**$60**
I would choose (please tick)		☐	☐	☐

Fig. 1 A sample of a choice task used in the survey

Table 1 Choice attributes and attribute levels with corresponding dummy-coding

Attribute	Level	Dummy Coding
Brown Kiwi (Native bird - flightless)	0 - Heard in 1 out of 200 planted forests	0,0 = current condition
	1 - Heard in 10 out of 200 planted forests	1,0 = intermediate level of increase
	2 - Heard in 20 out of 200 planted forests	0,1 = highest feasible level of increase
Giant Kokopu (Native fish)	0 - Seen in 1 out of 10 suitable streams	0,0
	1 - Seen in 3 out of 10 suitable streams	1,0
	2 - Seen in 5 out of 10 suitable streams	0,1
Kakabeak (Native plant)	0 - At least 3 naturally occurring shrubs	0,0
	1 - At least 10 actively managed shrubs	1,0
	2 - At least 20 actively managed	0,1

		shrubs	
Green gecko	0 -	Sighted in 1 out of 50 walks	0,0
(Native lizard)	1 -	Sighted in 3 out of 50 walks	
	2 -	Sighted in 5 out of 50 walks	1,0
			0,1
Bush Falcon	0 -	Sighted in 1 out of 8 drives	0,0
(Native bird – flyer)	1 -	Sighted in 3 out of 8 drives	
	2 -	Sighted in 5 out of 8 drives	1,0
			0,1
Price			0
($ per year for five			$30
years)			$60
			$90

2

3

Table 2 Sample distribution by choice task order and experimental design of the balanced sample

Choice Task Order	Number of Observed Choice Tasks			
	ORD	OOD	BDD	Pooled
1^{st}	56	56	56	168
2^{nd}	56	56	56	168
3^{rd}	55	55	55	165
4^{th}	56	56	56	168
5^{th}	56	56	56	168
6^{th}	56	56	56	168
7^{th}	56	56	56	168
8^{th}	56	56	56	168
9^{th}	56	56	56	168
Total choice observations	503	503	503	1509
Total number of respondents	57	59	56	172

Table 3 Evaluation of the statistical efficiency of the three designs

Statistical Efficiency Measure	Design Efficiency Values		
	ORD	BDD	OOD
Assuming βs = 0			
D_z-error	0.205	0.178	0.091
A_z-error	0.542	0.478	0.308
Assuming βs ≠ 0 but fixed			
D_p-error	0.290	0.213	0.589
A_p-error	0.801	0.595	3.417
Assuming βs ≠ 0 and accounting for uncertainty			
D_b-error	0.307	0.223	0.937
A_b-error	0.850	0.622	18.886

Table 4 Testing for the reliability of prior information from a pilot survey

Attribute	Pilot Sample (n=314)				Pooled Sample (n=1509)				% diff in WTP[a]
	Coeff.	Std Err	*p*-value	Marginal WTP_P	Coeff.	St. Error	*p*-value	Marginal WTP_F	
Brown kiwi 1	**0.462**	**0.252**	**0.07**	**$ 22.00**	**0.495**	**0.109**	**<0.01**	**$ 19.42**	**11.7%**
Brown kiwi 2	**0.591**	**0.251**	**0.02**	**$ 28.14**	**0.654**	**0.105**	**<0.01**	**$ 25.63**	**8.9%**
Giant kokopu 1	0.242	0.241	0.32	NS	**0.318**	**0.101**	**<0.01**	**$ 12.45**	**--**
Giant kokopu 2	0.286	0.248	0.25	NS	0.134	0.103	0.19	NS	**--**
Kakabeak 1	0.335	0.233	0.15	NS	0.179	0.103	0.08	NS	**--**
Kakabeak 2	0.112	0.251	0.66	NS	**0.228**	**0.103**	**0.03**	**$ 8.96**	**--**
Green gecko 1	0.190	0.246	0.44	NS	0.019	0.102	0.85	NS	**--**
Green gecko 2	**0.549**	**0.241**	**0.02**	**$ 26.14**	0.098	0.101	0.33	NS	**--**
Bush falcon 1	**0.550**	**0.253**	**0.03**	**$ 26.19**	**0.481**	**0.106**	**<0.01**	**$ 18.86**	**28.0%**
Bush falcon 2	**0.706**	**0.246**	**<0.01**	**$ 33.62**	**0.720**	**0.104**	**<0.01**	**$ 28.23**	**16.0%**
Cost to respondent	**-0.021**	**0.004**	**<0.01**	--	**-0.026**	**0.002**	**<0.01**	--	--
ASC for *status quo*	**0.876**	**0.413**	**0.03**		-0.159	0.171	0.35		
Pseudo-R^2	0.060				0.245				
Number of choice observations	314				1850				

[a] To calculate for the percentage difference in marginal WTP, we used the formula: %diff = $[(WTP_P - WTP_F)/ WTP_P] \times 100\%$

Note: NS means *not significant* at the 90% confidence level.

Table 5 Conditional logit model estimates for the three design criteria

Attribute	ORD Sample			BDD Sample			OOD Sample		
	Coeff.	Std Err	*p*-value	Coeff.	Std Err	*p*-value	Coeff.	Std Err	*p*-value
Brown kiwi 1	**0.471**	**0.209**	**0.02**	**0.377**	**0.179**	**0.04**	**0.606**	**0.198**	**<0.01**
Brown kiwi 2	**0.702**	**0.206**	**<0.01**	**0.456**	**0.168**	**0.01**	**0.749**	**0.191**	**<0.01**
Giant kokopu 1	**0.349**	**0.195**	**0.07**	**0.378**	**0.161**	**0.02**	0.164	0.180	0.36
Giant kokopu 2	0.242	0.202	0.23	-0.031	0.169	0.86	0.190	0.175	0.28
Kakabeak 1	0.259	0.185	0.16	-0.039	0.180	0.83	0.215	0.187	0.25
Kakabeak 2	-0.092	0.205	0.65	**0.436**	**0.165**	**0.01**	0.101	0.184	0.58
Green gecko 1	0.132	0.200	0.51	-0.053	0.167	0.75	-0.052	0.190	0.78
Green gecko 2	**0.443**	**0.197**	**0.03**	-0.179	0.167	0.29	0.135	0.180	0.45
Bush falcon 1	**0.499**	**0.208**	**0.02**	**0.567**	**0.170**	**<0.01**	0.290	0.196	0.14
Bush falcon 2	**0.823**	**0.202**	**<0.01**	**0.789**	**0.172**	**<0.01**	**0.549**	**0.186**	**<0.01**
Cost to respondent	**-0.026**	**0.003**	**<0.01**	**-0.020**	**0.003**	**<0.01**	**-0.032**	**0.003**	**<0.01**
ASC for *status quo*	**0.734**	**0.329**	**0.03**	-0.039	0.307	0.90	-0.378	0.273	0.17

Log-likelihood	-459.28	-497.66	-469.62
Pseudo R^2	0.169	0.099	0.150
Adjusted Pseudo R^2	0.147	0.078	0.128
Number of observations	503	503	503

Note: Figures in boldface font indicate statistically significant at the 90 percent confidence level.

Table 6a Percentage (%) of SR-ANA by attributes across design criteria

Attribute	ORD	BDD	OOD	Pooled
Brown kiwi	5.4	0.0	0.0	1.8
Giant kokopu	17.5	12.5	17.9	16.0
Kakabeak	14.3	10.7	14.3	13.1
Green gecko	17.5	7.2	7.2	10.6
Bush falcon	3.6	1.8	1.8	2.4
Average for all attributes	11.7	6.4	8.2	8.8
Minimum	3.6	0.0	0.0	1.8
Maximum	17.5	12.5	17.9	16.0

Table 6b SR-ANA (in %) by number of attributes ignored across design criteria

Number of attributes ignored	ORD	BDD	OOD	Pooled
0	68.2	73.2	69.6	70.3
1	14.3	23.3	21.5	19.7
2	13.9	1.8	7.2	7.6
3	0.0	1.8	1.8	1.2
4	1.8	0.0	0.0	0.6
5	1.8	0.0	0.0	0.6
Total	100.0	100.0	100.0	100.0

Table 7 Latent class model estimates for the three design treatments

Attribute	ORD Sample			BDD Sample			OOD Sample		
	$\hat{\beta}$	Std Err	*p*-value	$\hat{\beta}$	Std Err	*p*-value	$\hat{\beta}$	Std Err	*p*-value
Brown kiwi 1	**0.533**	**0.260**	**0.041**	**0.421**	**0.254**	**0.097**	**0.836**	**0.214**	**<0.001**
Brown kiwi 2	**0.985**	**0.277**	**<0.001**	**0.454**	**0.207**	**0.029**	**0.996**	**0.216**	**<0.001**
Native fish 1	0.141	0.323	0.660	0.235	0.230	0.310	0.140	0.226	0.540
Native fish 2	-0.469	0.314	0.140	-0.219	0.318	0.490	0.336	0.234	0.150
Native plant 1	-0.140	0.299	0.640	-0.106	0.301	0.730	**0.438**	**0.228**	**0.055**
Native plant 2	**-1.025**	**0.360**	**0.004**	**0.379**	**0.227**	**0.095**	0.214	0.229	0.350
Green gecko 1	**-1.035**	**0.374**	**0.006**	-0.365	0.244	0.130	0.017	0.238	0.940
Green gecko 2	**-0.571**	**0.316**	**0.071**	**-0.735**	**0.401**	**0.067**	0.108	0.224	0.630
Bush falcon 1	**0.636**	**0.260**	**0.015**	**0.599**	**0.211**	**0.005**	0.305	0.225	0.180
Bush falcon 2	**1.065**	**0.260**	**<0.001**	**0.791**	**0.221**	**<0.001**	**0.628**	**0.215**	**0.004**
Cost to respondent	**-0.090**	**0.008**	**<0.001**	**-0.067**	**0.007**	**<0.001**	**-0.139**	**0.016**	**<0.001**
ASC status quo	**-4.349**	**0.464**	**<0.001**	**-5.610**	**0.507**	**<0.001**	**-5.547**	**0.598**	**<0.001**

Latent Class (LC)	LC prob	R^2	LC prob	R^2	LC prob	R^2
LC1 - Full Attendance	0.010	0.433	0.236	0.430	0.219	0.630
LC2 - Ignored non-bird attributes	0.297	0.442	0.108	0.449	0.097	0.613
LC3 - Ignored SQ	0.454	0.038	0.364	0.018	0.319	0.000
LC4 - Ignored Cost	0.239	0.144	0.292	0.224	0.365	0.243
Total Prob/Overall R^2	*1.000*	*0.619*	*1.000*	*0.587*	*1.000*	*0.672*
Log-likelihood	-264.71		-305.01		-256.62	
BIC(LL)	590.06		670.39		574.40	
AIC(LL)	559.42		640.01		543.23	
AIC3(LL)	574.42		655.01		558.23	
Choice Observations	503		503		503	

Note: Text in boldface font indicates statistical significance at the 90 percent confidence level.

Appendix Table 1 Conditional logit model estimates using the pilot survey

Attribute	Coefficient	Std Err	*t*-ratio	*p*-value
Brown kiwi 1	**0.462**	**0.252**	**1.832**	**0.067**
Brown kiwi 2	**0.591**	**0.251**	**2.354**	**0.019**
Giant kokopu 1	0.242	0.241	1.002	0.316
Giant kokopu 2	0.286	0.248	1.155	0.248
Kakabeak 1	0.335	0.233	1.441	0.150
Kakabeak 2	0.112	0.251	0.446	0.655
Green gecko 1	0.190	0.246	0.771	0.441
Green gecko 2	**0.549**	**0.241**	**2.278**	**0.023**
Bush falcon 1	**0.550**	**0.253**	**2.174**	**0.030**
Bush falcon 2	**0.706**	**0.246**	**2.865**	**0.004**
Cost to respondent	**-0.021**	**0.004**	**-5.136**	**<0.001**
Indicator for SQ	**0.876**	**0.413**	**2.122**	**0.034**
Log-likelihood value				-324.473
Pseudo R^2				0.078
Adjusted Pseudo R^2				0.060
Number of choice observations				314
Number of respondents				35

Note: Text in boldface font indicates statistical significance at the 90 percent confidence level.

Appendix Table 2 Estimates of normalised AICs of panel latent class logit models using the three design samples

Specification	Latent classes (LCs) – Attributes ignored	Normalised AIC (AIC/N)		
		ORD	BDD	OOD
1	LC1 – Ignored SQ LC2 – Ignored non-bird attributes LC3 – Ignored all attributes	1.126	Did not converge	1.309
2	LC1 – Ignored SQ LC2 – Ignored Non-bird attributes LC3 – Ignored Cost	1.168	1.339	1.243
3	LC1 – Ignored SQ, LC2 – Ignored Non-bird attributes LC3 – Full attendance	1.135	1.362	1.309
4	LC1 – Ignored SQ LC2 – Ignored Non-bird attributes LC3 – Full attendance LC4 – Ignored all attributes	1.077	Did not converge	1.332

5	LC1 – Ignored cost	1.147	1.340	1.413
	LC2 – Ignored SQ			
	LC3 – Ignored Non-bird attributes			
	LC4 – Ignored all attributes			
6	LC1 – Ignored cost	1.172	1.342	1.085
	LC2 – Ignored SQ			
	LC3 – Ignored Non-bird attributes			
	LC4 – Ignored Falcon			
7	LC1 – Ignored cost	1.131	1.335	1.247
	LC2 – Ignored SQ			
	LC3 – Ignored Non-bird attributes			
	LC4 – Ignored Kiwi			
8	LC1 – Ignored SQ	1.139	1.365	Did not
	LC2 – Ignored Non-bird attributes			converge
	LC3 – Full attendance			
	LC4 – Ignored Kiwi			

9	LC1 – Ignored SQ	1.139	1.366	1.362
	LC2 – Ignored Non-bird attributes			
	LC3 – Full attendance			
	LC4 – Ignored Falcon			
10	LC1 - Ignored SQ	Did not	1.366	1.371
	LC2 – Ignored Non-bird attributes	converge		
	LC3 – Ignored Kiwi			
	LC4 – Ignored Falcon			
11	**LC1 – Full attendance**	**1.074**	**1.335**	**1.085**
	LC2 – Ignored SQ			
	LC3 – Ignored Non-bird attributes			
	LC4 – Ignored cost			

<u>Addressing the comments from the editor and reviewers</u> (EARE-D-12-00119R1**)**

Comments of editor and reviewers are in normal font, responses of authors are reported in bullet points in *italics.*

EARE-D-12-00119R1
June 5, 2014

Dear Richard Yao,

Thank you for submitting a revision of your paper, "Experimental Design Criteria and Their Behavioural Efficiency: An Evaluation in the Field" to Environmental & Resource Economics (ERE). I opted to send the paper out again for review, and now have heard back from the both of the original reviewers. The two reports are appended below.

I am pleased to say that both reviewers recommend acceptance of the paper subject to (a total of) three minor revisions.

- *Thank you very much for your message and comments.*

The one suggestion that warrants some thought is the request for some "discussion on weakness of using ANA as the measure of efficiency". I ask that you address this.

- *Thank you for pointing this out. A brief discussion on the weakness of ANA as the measure of efficiency is now written in Lines 66-75.*

In reading your paper closely I have a few comments and suggestions that I would like you to incorporate. One major concern I have had with this study is the sample size. Please be explicit in the text that your analysis is based on three subsamples of 56 respondents (unless I misunderstood something). Of course, even if all respondents were under the same experimental design, it is often difficult getting a choice experiment published with less than 200 respondents. The sample size does open up the criticism of whether your results are subject to sampling error as it could simply be by chance that there are correlations between the design and the presence of ANA. I am not suggesting you need to go out and collect more data. But instead just appropriately caveat the findings. On a related point, one is usually concerned with the typical estimators for the variance-covariance matrix when the number of independent observations is small. Does your analysis account for this?

- *We have now made it explicit in Lines 391-395 that we derived the 503 observations for each subsample from at least 56 respondents. We have now written that our total sample size was 172 respondents.*

- *To address your other concern, we have now written in Lines 394-397 that:*

> *"The pooled sample size of 172 would appear small if no allowance is made for the high efficiency of the designs used in this application. However, we note here that the asymptotic properties of the estimator converge at the unusual rate of the square root of the sample size and should already be effective at this number of respondents."*

Here are some minor suggestions:
1. Abstract. Delete the word “contributions”.

- *The word is now deleted.*

2. Abstract. Perhaps state instead “optimal orthogonal in the difference design” to be clearer. When I read “orthogonal design” and “optimal orthogonal design” I wondered how these could possibly be different (i.e. orthogonal designs are of course based on optimality criteria).

- *Thank you for this suggestion. We have now changed from “optimal orthogonality” to “optimal orthogonality in the difference” throughout the manuscript (e.g. Lines 7, 213). An orthogonal design is often not unique for a set of attributes and levels. The word “optimal” applies to the search for the most efficient of these orthogonal designs according to some a-priori and plausible assumption (e.g. the price coefficient should be negative, more is better, etc.).*

3. Introduction. A snapshot of CE applications is a lackluster way to begin this paper. I would simply delete this and begin by motivating the research with discussion of the need for assessing the efficiency of competing experimental designs.

- *Thank you for this suggestion. We have now deleted the snapshot and replaced it with the motivation of the research. Please see Lines 22 to 28.*

4. Page 2. I am not sure what you mean by “theoretically valid framework”. It would be hard to argue that all your respondents are in fact revealing their true preferences. I suppose it is valid conditional on respondents actually making choices that maximize utility.

- *Thank you for this suggestion. We have now deleted those words as those might confuse the readers.*

5. Page 3. Especially for the more casual reader, this discussion is not clear without at least a brief description of what you mean by serial ANA or the fully compensatory “assumption”.

- *Thank you for pointing this out. We now explain both serial non-attendance and fully compensatory choice behaviour. Please see Lines 52-58.*

6. Equation (6) should be reformatted as the lhs looks like D “minus” error.

- *Equation 6 now reformatted as suggested. Please see the row after Line 228.*

7. The mathematical notation is not consistent throughout, e.g., the beta vector is only sometimes bolded. I recommend bolding vectors and matrices throughout.

- *Thank you for pointing out this oversight. All vectors and matrices are now in boldface font throughout.*

8. First sentence of the conclusion: should be “design” rather than “designs”.

- *Thank you for this suggestion. We have now changed “designs” to “design”.*

9. The discussion on pages 16-17 was a bit difficult to follow. If I understand correctly, you use the stated assessments of ANA to define possible latent classes (e.g. a cost ANA class), but you do

not impose that a respondent that says they belong to a latent class to actually be in that class nor do you assign to them zero coefficients. Your approach makes sense, and avoids possible endogeneity concerns. But your discussion here can be condensed and what you do made more explicit. Perhaps place what others have done in a footnote.

- *You are correct, thank you for this suggestion. We have now rewritten Lines 331-345 accordingly and placed what others have done in Endnote number 4 (line 339), as suggested.*

10. Page 17, middle paragraph. Delete “though,”.

- *Thank you for this comment. “though” now deleted.*

At this point I am happy to recommend that your paper be accepted, conditional on addressing the remaining reviewer and editor comments. As I hope to simply accept your next revision “as is”, I ask you to make sure that the paper adheres to the ERE style guidelines and that you go over the paper carefully to correct any remaining grammatical errors.

- *Thank you for this suggestion. We have gone through the paper thoroughly and carefully corrected the minor grammatical errors and to our eyes it now fully adheres to the ERE style guidelines.*

Thank you again for your submission.

Best Regards,
Christian Vossler
Co-Editor, ERE

Reviewer #1: Some minor issues:

Update the reference
Hole A (2011) A discrete choice model with endogenous attribute attendance. Economic Letters, 110(3), 203-205

- *Thank you for this suggestion. Reference now updated accordingly.*

Page 2, line 25
Louviere and Woodworth (2003). It is 1983, not 2003

- *Thanks. “2003” now changed to “1983”.*

Reviewer #3: I appreciate the authors' responses and the improvement in clarity of the paper. I personally remain a bit skeptical of whether ANA is a "good" measure of behavioral efficiency (as opposed to a legitimate preference), but I agree with the author(s) that readers can make up their own mind and that some readers will agree and some will disagree. My only request is that you simply add some (small) discussion on weaknesses of using ANA as the measure of efficiency.

- *We have now elaborated on this (Lines 66-75) as requested. We have also added Endnote number 2 (Line 75) acknowledging and thanking an anonymous reviewer for this suggestion.*