UNIVERSITY OF WAIKATO

Hamilton
New Zealand

Designs efficiency for non-market valuation with choice modelling: how to measure it, what to report and why

John M. Rose

Riccardo Scarpa

Department of Economics

Working Paper in Economics 21/07

October 2007

John M. Rose
Institute for Transport and Logistics Studies
University of Sydney
144 Burren St.
Sydney, New South Wales

Tel: +61 (0)2 9351 3076
Fax: +61 (0)2 9351 4433

Email: johnr@itls.usyd.edu.au
Web: http://www.itls.usyd.edu.au

Riccardo Scarpa
Economics Department
University of Waikato
Private Bag 3105
Hamilton, New Zealand

Tel: +64 (0) 7-838-4045
Fax: +64 (0) 7-838-4331

Email: rscarpa@waikato.ac.nz
Web: http://www.mngt.waikato.ac.nz

Abstract

We review the basic principles for the evaluation of design efficiency in discrete choice modelling with a focus on efficiency of WTP estimates from the multinomial logit model. The discussion is developed under the realistic assumption that researchers can plausibly define a prior on the utility coefficients. Some new measures of design performance in applied studies are proposed and their rationale discussed. An empirical example based on the generation and comparison of fifteen separate designs from a common set of assumptions illustrates the relevant considerations to the context of non-market valuation, with particular emphasis placed on C-efficiency. Conclusions are drawn for the practice of reporting in non-market valuation and for future work on design research.

Keywords
Efficient experimental design
Multinomial Logit
Random utility model
Choice modeling
Nonmarket valuation
C-efficiency

JEL Classification
C25; Q51

Acknowledgements
We thankfully acknowledge competent research assistance from Andrew Collins. The usual disclaimer on the remaining errors applies.

1. Introduction

Stated choice modelling has now an established role in nonmarket valuation. Practitioners are engaged in testing the method and defining the boundaries of its use in public decision making and cost benefit analysis. In this respect the method has taken up a research agenda which is quite distinctive from other fields of applications, such as in transport, marketing, food choice and health research. One of the areas of distinctiveness is associated with the methodology of experimental design for the specific purpose of deriving nonmarket values.

A survey of existing nonmarket valuation studies indicates that there is a prevailing format of stated choice surveys in nonmarket valuation. Typically, this involves asking respondents to indicate their preferred alternative from those offered within a given choice set. Alternatives in the choice set are often outcomes of policies that can vary in their effects of relevance to the respondent. Effects of policies are described by a selected number of attributes, each of which can take a qualitative or numerical level. Rather than review a single choice set, respondents are typically asked to evaluate several, thus increasing the number of observations per individual surveyed. Experimental design is used to allocate levels to attributes of alternatives in choice sets. As such, experimental designs lie at the core of all stated choice studies. Conceptually, experimental designs may be viewed as the systematic arrangement in matrices of the values that researchers use to describe the attributes representing the alternative policy options of the *hypothetical* choice situations. Because the combinations of attribute and attribute levels can be huge even with relatively simple problems, some theory must be used to drive the selection of these levels and their arrangements in choice sets in order to achieve the required information within practical sample sizes.

Via experimental design theory, the analyst is able to determine the values to be assigned to attributes in each alternative situated within the choice sets to be used in the survey. The assignment of these values occurs in some systematic (i.e., non-random) manner so as to achieve the intended results of the survey in an efficient, i.e., a least cost manner. The theory makes use of various criteria to evaluate the outcomes of these assignments on the basis of the assumptions invoked by the analyst as incorporated by a given model specification. The selection of the correct set of criteria will drive the analyst to an adequate choice for the purpose at hand and conditional on the chosen specification and other assumptions made by the researcher.

Experimental design techniques are of general relevance in survey research. However, the specific focus of nonmarket valuation on the derivation of implicit prices from discrete choices has some important and distinctive implications in the practice of experimental design which are still inadequately addressed, as discussed in depth by Ferrini and Scarpa (2007). The present paper intends to contribute to developing an understanding of these implications within the 'workhorse' of discrete choice analysis: the conditional logit model predicated on random utility theory. Extensions to other specifications of the logit family are conceptually immediate, although technically challenging, and definitely beyond the scope of this paper.

To do so we selectively draw from the wide and rapidly expanding literature in experimental design for logit models and we propose an infrequently used criterion based on the specific needs of nonmarket valuation. For CM surveys developed to estimate monetary values desirable criteria should revolve around efficiency of willingness to pay (WTP) estimates, which are functions of the utility parameter estimates of logit models predicated on random utility theory. While criteria measuring predictive performance of probabilities, utility balance across alternatives and efficiency of the utility estimates are much more frequently used in design evaluation, the way such criteria are related to efficiency of and sample size requirements for WTP estimates is unclear. In this paper we set up the building blocks for investigating such a relationship and provide a worked out example exploring the relationship between parameter efficiency and WTP efficiency. We set-up our example in a setting that is most common in nonmarket valuation applications, the one with repeated choices from two hypothetical alternative and the status-quo or no-buy option.

The rest of the paper is organised as follows. Section 2 provides a discussion of the relationship between discrete choice models, random utility theory and experimental design. Section 3 discusses various efficiency criteria that have been employed in the literature before Section 4 introduces a new criteria based on WTP efficiency. Section 5 provides a brief discussion on what should be reported in terms of statistical measures after which algorithms for generating efficient designs are introduced. Section 7 provides a treatise on the issue of scaling and designs which has often been ignored within the literature. A case study in which 15 different experimental designs are generated using different design strategies is next presented, before general conclusions are made.

2. Discrete choice models, random utility and experimental design

Qualitative choice is based on discrete outcomes represented by the selection of alternatives from a given consideration set. What form of evaluation (lexicographic, elimination by aspect, economic or other attribute screening rules, etc.) is predominant amongst respondents in driving such selection, remains an elusive issue. Much research is being conducted on methods to practically distinguish these processes starting from observed behaviour. Regardless of actual evaluation processes, in applied research the most successful paradigm to date has been random utility theory (RUT), and we refer to this in what follows. Similarly, in terms of statistical analysis of responses, the most successful specification consistent with RUT has been the conditional logit model (McFadden 1974). This model remains at the core of most of the more sophisticated specifications, such as nested and mixed logit models. What is discussed and illustrated in practice here can be easily extended, although not so easily illustrated, to more sophisticated RUT-based models.

The main point of departure of our study concerns the logical consequences from being able to assume the direction and sometime the relative magnitude of the values of the taste intensity parameters in the utility function. As soon as the researcher can plausibly defend that some attributes of choice may generally be expected to have a given sign or relative size the efficiency of the design for a logit specification can easily be shown to be improved from what would be the case in the absence of such an assumption. In this respect our work cannot be compared to similar research

carried out within the limited framework of probability balanced designs, that are predicated on researchers' ignorance of the values of taste intensities (Burgess and Street 2005; Street and Burgess 2005; Street et al. 2001, 2005). In our case then, we take a completely opposite approach from the stance taken by Lusk and Norwood (2005), who state that:

> "…in many cases researchers do not have strong priors regarding preferences. This article focuses on design strategies where the analyst has no prior information about true utility." (Lusk and Norwood, AJAE 2005(97(3):772))

With this premise, the authors proceed to develop a discussion prevalently based on the property of orthogonality, which is—as they themselves note—much more relevant for designs developed for linear multivariate models than it is for highly non-linear models such as those in the logit family.

As a matter of fact, we argue exactly the opposite, which is that in the greatest majority of nonmarket valuation studies researchers indeed *are* able to predict at least the sign of the price coefficient. In reality, however, researchers can normally do more than this and express some beliefs on the range of values that are likely to be taken up by other parameters in the utility function.

In terms of assumptions our research is therefore more akin to research efforts by Sandor and Wedel (2001, 2002, 2005), Bliemer and Rose (2005), Bliemer et al. (2005, 2007) Ferrini and Scarpa (2007) and Kessels et al. (2006). We also note that this approach is more in keeping with previous literature in optimal design for non-market valuation (Alberini 1995, Kanninen 1993a, b), and of sequential improvement of survey designs in non-market valuation (Kanninen 1993b; Scarpa et al. 2007).

We will show with examples that when adequately expressed this a-priori information is of great use and can lead to substantial efficiency in the design. In doing so, however, the analyst must be made aware of some potential difficulties, some of which are of specific interest to the current choice modeling practice for the purpose of non-market valuation, such as the effect of the status-quo alternative and that of the choice of attribute coding on the evaluation of the efficiency of the design.

We now move our attention to the definition of efficiency in the context of the logit model commonly used to derive estimates of utility coefficients from observed discrete choice.

3. Measuring design efficiency for taste intensities

In this section we examine the measures of design efficiency that are of interest when the objective is to estimate the coefficients of the indirect utility function, or the so called taste intensities.

3.1 The basics

Consider a situation involving the choice between j=1,2,…,J alternatives, each of which are described by K attributes. Assuming the choice process is modeled using a conditional logit specification with Gumbel error scale $\lambda > 0$, we get:

$$\Pr(Y_i = j) = \frac{e^{\lambda\beta' x_{ij}}}{\sum_{j=1}^{J} e^{\lambda\beta' x_{ij}}}, \lambda > 0, \qquad (1)$$

which is the probability that alternative j will be selected in choice task i.

The specific values of x_{ij} are defined by the experimental design. An efficient design will minimize the variance-covariance estimator, or—put differently—will maximize the amount of information the design conveys to identify the estimates of the vector β. The information matrix for the design under the conditional logit assumption is given by the matrix of second derivatives of the log-likelihood function, which can compactly be written as:

$$I(\beta, x_{ij}) = \frac{\partial^2 \ln L}{\partial\beta\partial\beta'} = -\sum_{i=1}^{n}\sum_{j=1}^{J} P_{ij}(x_{ij} - \bar{x}_{ij})(x_{ij} - \bar{x}_{ij})', \quad \text{with } \bar{x}_{ij} \equiv \sum_{j=1}^{J} P_{ij} x_{ij}, \quad (2)$$

which is a matrix of size $K\times K$.

One of the reasons of the popularity of the multinomial logit model is that of having a relatively simple mathematical formulation of both the Jacobian (gradient or vector of first derivatives) and Hessian (matrix of second derivatives). Both objects however, are functions of both the utility coefficients β and the matrix of choice attributes x_{ij} (i.e., the experimental design). So, an informative design is one that makes some function of the size of $I(\beta,x_{ij})$. In other words, taken $g(I(\beta,x_{ij}))$ as a measure of information, an informative design should make this measure large. At this stage it is useful to revise the relationship between $I(\beta,x_{ij})$ and a common Maximum Likelihood estimator of the asymptotic variance-covariance (AVC) matrix $\Sigma(\beta,x_{ij})$ of a design. The Maximum Likelihood estimator of the AVC matrix for a design to be used with the conditional logit model is the negative inverse of the expected Fisher information matrix (e.g., see Train 2003), where the latter is equal to the second derivatives of the log-likelihood function:

$$AVC = \Sigma(\beta, x_{ij}) = \left[E\left[I(\beta, x_{ij})\right]\right]^{-1} = \left[-\frac{\partial^2 \ln L}{\partial\beta\partial\beta'}\right]^{-1}, \qquad (3)$$

where $\ln L$ is the log-likelihood of the design:

$$\ln L = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} \ln P_{ij}(x_{ij}, \beta), \qquad (4)$$

where i choice tasks implied in the design, and j the alternatives.

In choosing an informative (efficient) design that one can choose to think in equivalent terms of either *maximizing* information or *minimizing* variance. A suitable algorithm would search the arrangement of attribute and levels in a suitably coded matrix x_{ij} such that an optimal solution is found according to some stopping criteria.

3.2 Design efficiency measures

A key passage is the definition of the function $g(\cdot)$, which is useful to define as a single number, rather than a collection as in vectors and matrices. A convenient scalar measure of the size of a matrix is its determinant, which is a sum of terms each made-up of products of systematically selected elements of the matrix. A nonzero determinant matrix implies the matrix has full rank (no collinearity and identification of the β). So, the determinant of the information matrix (or equivalently minimizing that of the AVC) is a valid measure of efficiency of a candidate design. However, the determinant will be larger as k—the number of elements in β—increases, so that one must devise a measure that accounts for that too. An often used measure is the D-error:

$$D_p error = \det\left(\Omega(\beta, x_{ij})\right)^{1/k} \qquad (5)$$

So, that a search over the arrangement of attribute levels in x_{ij} can be used to minimize such scalar measure.

Rather than the determinant, another measure of efficiency for taste intensities has been used (e.g., Louviere et al. 2003) the so called *A*-efficiency, defined as the trace of the AVC:

$$A = trace\left(\Omega(\beta, x_{ij})\right). \qquad (6)$$

However, this measure seems to have encountered lower acceptance and use within the published literature.

One final measure, which we explore in this paper, does not look at the AVC matrix, but rather the choice probabilities for the design. This measure, proposed by Kessels et al. (2006), is not explicitly meant to be used as a measure of design efficiency, however we use it here as a means of attempting to remove alternatives that may be dominated. The probability or utility balance of a design is given as:

$$B = \frac{\sum_{s=1}^{S}\left(\prod_{j=1}^{J}\Pr(Y_i = j)\right)}{S\left(\frac{1}{J}\right)^{J}}. \qquad (7)$$

Eq. (7) will range between zero and 100 percent, with the percentage value representing how balanced the probabilities (or utilities) are over the alternatives within the design. A zero value indicates that there exists a completely dominate

alternative within each choice set, whereas a value of 100 percent indicates that each alternative in every choice set has an equal probability of being chosen.
We note that although we deliberately restricted the discussion to the conditional logit model, the principles are fully applicable to any model of discrete choice, such as the nested logit or mixed logit models. All it requires is the computation of the adequate information matrix (see e.g., Bliemer and Rose 2006).

3.3 Design specificity and coefficient uncertainty

Two important observations are in order here, both of which clearly affect the measurement of efficiency of a conditional logit design. The first concerns the coding of the variables in the matrix x_{ij} and it concerns the fact that efficiency depends on the type of coding chosen (on the levels, effect-coding, or dummy variable coding). As a consequence, a design obtained under effect-coding will have different efficiency value if the coding applied in estimation is dummy variable coding. Hence the efficiency measures should not be compared across models with different coding applied to the same design.

The second concerns the assumptions about the values of β, which are the very quest of a stated choice survey study and hence cannot be known with certainty at the time of designing the experiment. However, they can be assumed with uncertainty by the analyst and such uncertainty can be formally defined in terms of a-priori distributions.

For this reason the literature distinguish between point D-error and Bayesian D-errors, using the notation D_p and D_b respectively. The latter is just an expectation taken over the assumed a-priori distributions of β. Suppose, for example that the values of β are a priori believed to be distributed Normally, with a vector of means μ and matrix of variance Σ, then the D_b error would be:

$$D_b\ error = \int \left[\det\left(\Omega(\beta, x_{ij})\right) \right]^{1/k} N(\mu, \Sigma) d\beta \qquad (8)$$

Of course, less informative priors can be invoked, such as uniform distributions over a broad range of values.

3.4 Level of significance and design replicates

Of course, typically the survey will produce many replications of the same design as generally a design will be completed by more than one respondent. In generating the design, it is common to assume only a single respondent, however, this need not always be the case. In particular, it is useful to assume more than one respondent when subsets of a design are to be given to different respondents (this is commonly achieved, e.g., via *blocking* of the design). Suppose that a design is broken into *G* subsets, with *n* respondents reviewing each subset (noting that *n* may be different for each *G*). Then the AVC matrix for the final model would be:

$$\Omega_N(\beta, x_{ij}) = \sum_{g_n=1}^{G_n} \Omega_{g_n}(\beta, x_{ij}) \qquad (9)$$

Further, note that the AVC has dimension $K \times K$ and that the asymptotic standard errors for each estimate of the elements of β are given by the squared root of the diagonal of the AVC matrix:

$$\begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_k \end{bmatrix} = \sqrt{diag\left(\Sigma(\beta, x_{ij})\right)} \qquad (10)$$

This is sometimes used to derive a measure of the (theoretically) required design replicates to achieve a given significance value for a choice attribute coefficient *k* via the required *t*-value and the relationship:

$$t_{\beta_k} = \frac{\beta_k}{s_k}\sqrt{n_{\beta_k}} \rightarrow n_{\beta_k} = \left[\frac{t_{\beta_k} s_k}{\beta_k}\right]^2 \qquad (11)$$

For example, suppose one assumes a $\beta_1 = 1.2$ and derives a design with an $s_1 = 2$ but wants to compute the number of design replicates necessary to achieve a five percent significance for which the two-tailed *t*-value is ≈1.96. Then an adequate design size can be of 11 replicates since:

$$n_{\beta_k} = \left[\frac{t_{\beta_k} s_k}{\beta_k}\right]^2 = \left[\frac{1.96 \times 2}{1.2}\right]^2 = 10.67 \approx 11 \qquad (12)$$

If the design is segmented into three different subsets consisting of different choice sets, than one would need about 32-33 respondents to achieve five percent significance, assuming the prior parameter is correct. Such a calculation may be made for all *k* parameters, with the theoretical minimum sample size being the largest value calculated (see e.g., Bliemer and Rose (2005); designs that seek to minimize the sample size are termed S-efficient designs). Although this illustration is informative to clarify the relationship between design and sample size required to achieve significance of β estimates, this is obviously a theoretical relationship and the selected model is only a simplification of the real world, so that typically larger sample sizes are necessary than those indicated. How much larger will depend on the empirical case at hand.

4. Design efficiency for prediction and for WTP

In many marketing and transport studies choice experiments are used to derive predictions of choices, and in particular predictions on the effect of changes in the choice attributes. So, other criteria rather than efficiency need be used to assess designs when the stated choice exercise has this purpose. Kessels et al. (2006) propose the use of G- and V- optimality criteria for the experimental choice context. These criteria measure the variance of prediction, rather than the variance of the taste intensities. In particular, G-optimality relates to the minimization of the *maximum*

prediction variance in the design, while V-optimality relates to the minimization of the *average* prediction variance.

Finally, of central interest to the literature in non-market valuation is the concept of C-optimality, first introduced in the literature by Kanninen (1993a,b). This criterion is specifically suited for minimizing the variance of functions of model coefficient estimates, such as willingness to pay. A frequently adopted specification for indirect utility is linear in the parameter and specified over choice attributes, one of which, for valuation studies, must be the cost of the alternative. In these context, it can be shown that the unit WTP for the attribute can be derived as a function of the coefficient attributes:

$$WTP_k = \frac{\beta_k}{-\beta_{\text{cost}}} \tag{13}$$

This is a highly nonlinear function of the coefficient estimates and the variance of this can be approximated using the delta method.

The ML estimator for β is asymptotically normal, so that given consistency:

$$\sqrt{n}\left(\beta_{ML} - \beta\right) \xrightarrow{N} N(0, Var(\beta_{ML})) \tag{14}$$

Take any continuous function twice or more differentiable g(β). Use the first two terms of a Taylor series approximation to expand it around the estimates as follows:

$$g\left(\beta_{ML}\right) \approx g(\beta) + \nabla g(\beta)'(\beta_{ML} - \beta) \tag{15}$$

Where $\nabla g(\beta)$ is the vector of first derivative (gradient of g()) and ' indicates transposition.

We can compute the variance of this linear function so that:

$$Var\left[g\left(\beta_{ML}\right)\right] \approx \nabla g(\beta)' Var(\beta_{ML}) \nabla g(\beta). \tag{16}$$

Having this approximation all we need to do now is to substitute g(.) with -α/β, where to avoid notational clutter induced by the use of sub-scripts we indicate with α the taste intensity of the generic attribute and with β the cost coefficient.

First note that, $\frac{\alpha}{-\beta} = -\alpha(\beta)^{-1}$, this makes the use of the product rule to derive the gradient easier:

$$\nabla g(-\alpha\beta^{-1}) = \begin{bmatrix} f' \\ h' \end{bmatrix} = \begin{bmatrix} \frac{\partial(-\alpha\beta^{-1})}{\partial\alpha} \\ \frac{\partial(-\alpha\beta^{-1})}{\partial\beta} \end{bmatrix} = \begin{bmatrix} -\beta^{-1} \\ \alpha\beta^{-2} \end{bmatrix} \tag{17}$$

So that:

$$Var\left[g\left(\beta_{ML}\right)\right] \approx \nabla g(\beta)' Var(\beta_{ML}) \nabla g(\beta) =$$
$$\left[-\beta^{-1} \quad \alpha\beta^{-2}\right]\begin{bmatrix} Var(\alpha) & Cov(\alpha,\beta) \\ Cov(\alpha,\beta) & Var(\beta) \end{bmatrix}\begin{bmatrix} -\beta^{-1} \\ \alpha\beta^{-2} \end{bmatrix} \tag{18}$$

Multiplying the first row vector by the matrix gives:

$$\left[-\beta^{-1}V(\alpha) + \alpha\beta^{-2}C(\alpha,\beta) \quad -\beta^{-1}C(\alpha,\beta) + \alpha\beta^{-2}V(\beta)\right] \tag{19}$$

Then, multiplying the resulting row vector by the final column vector gives:

$$-\beta^{-1}\left[-\beta^{-1}V(\alpha) + \alpha\beta^{-2}C(\alpha,\beta)\right] + \alpha\beta^{-2}\left[-\beta^{-1}C(\alpha,\beta) + \alpha\beta^{-2}V(\beta)\right]$$
$$\rightarrow Var\left(\frac{\alpha}{-\beta}\right) \cong \beta^{-2}\left[V(\alpha) - 2\alpha\beta^{-1}C(\alpha,\beta) + (\alpha/\beta)^2 V(\beta)\right] \tag{20}$$

So, the C-criterion relates to the minimization of such variance. One thing to note is that, unlike in the case of CVM in which there is only one WTP to derive, here the variance relates to an element of *k*-1 WTPs. Furthermore, different attributes may be described in different units. So, for example, with an attribute expressed in miles and one in number of properties affected the WTP per unit will be referring to different measures. Suppose one takes the sum of the *k*-1 variances, then minimizing such sum may result in an unsatisfactory outcome if the minimum is obtained by diminishing the variance unevenly across WTPs. For example, the minimum may be reached by achieving a very small variance for attribute 1 while leaving the variance for attribute 2 higher than desirable. Eq. (12) suggests a potential criterion, which is that of either maximizing the minimum *t*-value for the WTP:

$$x_{ij}^{*} = \arg\max_{x_{ij}}\left(\min\begin{bmatrix} t_{WTP_1} \\ \vdots \\ t_{WTP_{k-1}} \end{bmatrix}\right) \tag{21}$$

or equivalently, that of minimizing the number of design replicates necessary to achieve the desired significance level for WTP:

$$x_{ij}^{*} = \arg\min_{x_{ij}}\left(\max\begin{bmatrix} D_{WTP_1} \\ \vdots \\ D_{WTP_{k-1}} \end{bmatrix}\right) \tag{22}$$

To our knowledge neither of these criteria has been used so far in the literature of choice experiment design. We note in passing that all these criteria can be adapted so as to be amenable to a Bayesian prior as discussed in section 3.3.

In conclusion of this criteria review we emphasize how various criteria are available to evaluate a candidate design and each is particularly suitable to a specific purpose. Of course when the stated choice exercise has a variety if purposes, then perhaps a

weighted combination of selected criteria can be employed to derive the optimal design x_{ij}^{*}. A similar observation can be extended to the final specification. If the data collection is likely to support a variety of specifications, then the AVC matrix may be substituted by an adequate mixture of AVC matrix, one for each specification. However, we do not venture our empirical illustration in this territory, but note that could constitute fertile ground for further research.

5. What design efficiency measure to report?

Ideally one would like to know exactly what the true model is in terms of both specification and β values. Of course this is not attainable in practice. If it were, one would have an ideal measure against which to gauge the particular design used in the study. Nevertheless, on the basis of what has been discussed thus far, we are able to make meaningful recommendations on what statistics to report in a study with regards to the particular design employed. One must realize that there are two separate moments in a study. An initial stage at which one can plausibly postulate some prior for β on the basis of theory (e.g., the β for cost is negative and β for something good—such as clean air—is positive), and formalize these expectation via a distribution over a range of values. A final end-of-study stage at which one has the sample in hand and can derive an estimate of the population parameters conditional on the collected data. These estimates are the best available at that stage, and might be quite different from those postulated at the initial stage. The *true* values of the population coefficients, however, are still uncertain. So, although an absolute efficiency ratio cannot be provided, one can compute the relative efficiency of the initial stage design using as a bench mark the end of study design. Denoting by the superscript 0 the initial stage priors and with 1 the end of study estimates we recommend to report:

$$\frac{F\left(\hat{\beta}^{0}, x_{ij}\right)}{F\left(\hat{\beta}^{1}, x_{ij}^{*}\right)}, \tag{23}$$

where F denotes the particular criterion of interest, and the starred design indicates optimization with respect to the end of study estimates. We would argue that any other measure is not only relatively uninformative, but in some cases it can even be misguiding. Consider the frequent practice of reporting $100 \times N|X'X|^{1/k}$ (e.g., Lusk and Norwood 2005, p772) where N is the number of observation in the design and X the generic design matrix. This measure is virtually irrelevant with respect to the operating conditions of discrete choice modelling under random utility models. Additional criteria might also be reported to understand the relationship between the designs employed—which presumably has been derived by optimizing according to some valid criterion—and the values that the same design affords with regards to other criteria. So, for example, suppose one has obtained the design x_{ij}^{V} used in the study by optimizing for the *V-p* criterion, then it would probably be of interest to contrast this design by using the more common D-p criterion:

$$\frac{D\left(\hat{\beta}^{0}, x_{ij}^{V}\right)}{D\left(\hat{\beta}^{1}, x_{ij}^{*}\right)} \tag{24}$$

A high value of this ratio would illustrate that despite having been derived with a criterion that maximized efficiency in prediction, it turned out to perform well relative to a design that would have been optimized for efficiency in coefficient estimates.

6. Algorithms for design optimization for efficient designs

We now turn our attention to a brief description of the various algorithms proposed in the literature to search for improvements on a basic starting design, which can be—for example—the typical fractional orthogonal of the full factorial. Unfortunately, there does not exist much theoretical guidance as to which method should be employed. We are also not aware of studies that tested which type of design construction method is likely to produce the best results under various circumstances in practice. A number of algorithms have been proposed and implemented within the literature to systematically search the various attribute level arrangements to identify efficient designs. These algorithms operate mostly by systematically operating swaps across the rows and columns of the matrix x_{ij}. Typically, algorithms fall into one of two categories; row and column based algorithms.

In *row based algorithms*, a large number of choice sets are first generated from which choice sets to be used in the survey are selected. Typically, the choice sets are drawn from a full factorial design, although in many instances the full factorial will be too large (even with today's computing power) and fractional factorials may be generated instead. This is precisely what the most widely used row based algorithm, the *Modified Federov algorithm* (Cook and Nachtsheim 1980), does. The algorithm randomly draws *s* choice sets from either a full factorial or fractional factorial design, with the D-error of each random selection being calculated. The combination of choice sets that produce the lowest D-error is retained as the most efficient design. The algorithm is terminated either manually by the researcher, when some stopping criteria is achieved (e.g., no improvement in the D-error is achieved for 30 minutes) or when all possible choice set combinations has been explored. Row based algorithms have the advantage of being able to reject poor choice set candidates at the initial stage (e.g., choice sets in which the attributes of one or more alternatives are dominated or where a particular combination of attributes realistically cannot exist), and as such, these choice sets will never appear in the final survey. Nevertheless, row based algorithms generally find it difficult to maintain attribute level balance (where each attribute level appears an equal number of times over the design).

Column based algorithms on the other hand, begin by randomly generating a design and then systematically change the levels within each column (representing an attribute in the survey) of the design. Whilst it is difficult to reject poor choice sets using column based algorithms, such algorithms typically are able to maintain attribute level balance, particularly if the initially generated design has such a property. In general, column based algorithms offer more flexibility and are generally easier to use when dealing with designs with many choice situations, but in some cases (e.g., for unlabeled choice experiments and for specific designs such as those where certain attribute level combinations are forbidden) row based algorithms may be more suitable.

Rather than relying solely on row based or column based algorithms, some authors suggest using combinations of both. Huber and Zwerina (1996) implemented the RSC algorithm (*Relabeling*, *Swapping* and *Cycling*), which remains the most widely used algorithm today. The RSC algorithm alternates between *relabeling* (column based), *swapping* (column based), and *cycling* (row based) over many iterations. During the *relabeling* phase, all occurrences for two or more attribute levels within a column of the design are switched (e.g., if attribute levels 1 and 4 are relabelled then the column containing the sequence of levels {1,3,4,2,4,1,3,2} would become {4,3,1,2,1,4,3,2}). The *swapping* phase of the algorithm is similar to that of relabeling, however only a few of the attribute levels are changed within the column (e.g., swapping the first and third values in {1,3,4,2,4,1,3,2} would yield {4,3,1,2,4,1,3,2}). The *cycling* phase of the algorithm is row based, where the attribute levels are switched (similar to relabeling but now across rows, not down columns) within choice sets, one choice set at a time. The algorithm will generally try a number of iterations of either *relabeling*, *swapping* or *cycling*, before switching to another phase (typically randomly). Note that not all phases have to be used with various combinations of RSC being possible.

7. The impact of scale on willingness to pay

One consideration must be made at this stage about the scale parameter, which is often a neglected issue in D-efficient designs. This is particularly relevant when the focus is on WTP estimation and when a status-quo constant (or any alternative-specific constant) is expected to be part of the utility function. WTP is a one-to-many mapping of the vector β. In fact, infinite pairs of β_{-p} (non-price coefficients) and β_p produce the same vector of WTP values. Suppose, the values of β are as assumed above. Scaling them all by any positive constant produces the same WTP estimates. So implicitly in the assumption of values for β there is an assumption of the scale coefficient.

When—instead—utility includes an alternative-specific constant of some sort, scaling the vector β by any amount has an effect on the utility differences across alternatives, which are not scaled by the same constant. So, depending on the assumed scale of the Gumbel error, the same WTP vector can be associated with large or small utility differences with the status-quo, and hence different choice probabilities. Table 1 illustrates this case in which the levels of the attributes in the SQ choice are assumed to be the baseline (equal to zero) and hence the levels in the designed alternatives 1 and 2 are expressed as differences from those in the SQ.

This is, of course a corollary to the fact that with a high scale (small error variance) the choice probabilities become deterministic. However, it highlights how important an adequate specification of the error scale is to the evaluation of the design in the presence of alternative-specific constants. For a given scale though, the criteria of different designs can be compared. We hence now turn to a comparison of designs generated under the assumption of a multinomial logit specification for a given case study.

Table 1: Demonstration of impact of scaling on model outcomes

				λ = 1				
β	**1**	**1**	**2**	**3**	**-1**	**β'x_j**	**Δ$V_{j\text{-}sq}$**	**Pr(*j*)**
x_1	0	1	2	2	2	9	8	0.952
x_2	0	2	2	1	3	6	5	0.047
sq	1	0	0	0	0	1	0	0.000
WTP	1	1	2	3	-1			
				λ = 0.5				
β	**0.5**	**0.5**	**1**	**1.5**	**-0.5**	**β'x_j**	**Δ$V_{j\text{-}sq}$**	**Pr(*j*)**
x_1	0	1	2	2	2	4.5	4	0.806
x_2	0	2	2	1	3	3	2.5	0.180
sq	1	0	0	0	0	0.5	0	0.015
WTP	1	1	2	3	-1			
				λ = 0.2				
β	**0.2**	**0.2**	**0.4**	**0.6**	**-0.2**	**β'x_j**	**Δ$V_{j\text{-}sq}$**	**Pr(*j*)**
x_1	0	1	2	2	2	1.8	1.6	0.571
x_2	0	2	2	1	3	1.2	1	0.313
sq	1	0	0	0	0	0.2	0	0.115
WTP	1	1	2	3	-1			

8. Case Study

8.1 The Case Study Setting

This case study is devised to illustrate the considerations a researcher can make when engaged in developing a "typical" non-market valuation study. A recent review on the design solutions used in published non-market valuation studies (Ferrini and Scarpa 2007) suggests that a common set-up is that of what Louviere et al. (2000) called an unlabelled design based on a choice task involving the indication of the favourite alternative amongst three. Two of these have levels and attributes developed on the basis of a design, while the third represents the status-quo (see Breffle and Rowe (2002) for a discussion of the inclusion of the status-quo alternatives in non-market valuation studies and Scarpa et al. (2005) for some econometric insights). We hence adopt this framework, but caution the reader that generalizing the results from this case study to other contexts might well be unwarranted.

Most published studies investigate a range of 3-6 choice attributes plus the cost of the package to the respondent. We hence present results of a design with three attributes plus price and a status-quo constant. We postulate that the analyst is able to define some a-priori beliefs on the values of the β vector that can be adequately formalized. We assume that since much of the literature reports positive status-quo effects, the element of β relating to the status-quo is assumed to be positive and equal to unity. The price effect is of course negative and also equal to one. The three attributes differentiating the alternatives are assumed to be expressed as positive effects on utility and orderable in terms of a gradient one, two, and three. While one can very frequently express attributes in a way that can be generally expected to be perceived and evaluated by respondents as having a specific directional (positive or negative) effect on utility, the cardinal scaling is arguably the strongest a-priori assumption. However, this assumption can be relaxed by assuming a distributional form with overlapping densities, as we will see later.

The size of the design is of 20 choice sets, and the design attributes can all take four values (0,1,2,3) except price which can take five levels (these are 0,1,2,3,4). A size of 20 is not unusual and can be shared out across five, four or two respondents to obtain a balanced panel of respectively four, five and 10 choices per respondent. In non-market valuation studies it is frequently found that the number of levels used for the price attribute is larger than those used for non-price attributes.

8.2 Design Procedure Exploration

Fifteen designs are generated and compared across a range of criteria. In order to demonstrate why it is important to use experimental designs for stated preference studies, the first two designs we report were constructed using a purely random allocation of the attribute levels to the design. In generating the first design, we do not assume attribute level balance (i.e., each level of an attribute may appear an uneven number of times over the 20 choice sets), whereas for the second design, attribute level balance was enforced as a design criteria. All remaining designs also assume attribute level balance. Unlike Designs 1 and 2, Designs 3 to 5 and 9 to 15 were constructed using the RSC algorithm (see Section 5) assuming (different) optimisation criterion. Design 6 was constructed in a manner for which the RSC algorithm was not appropriate and hence only swapping was used. Designs 7 and 8 are orthogonal designs, for which the RSC algorithm is also inappropriate.

Designs 3, 4 and 5 represent designs constructing using the D-efficient criteria given as Eq. (5), and they illustrate the effect of varying the scale parameter in this context, as discussed in Section 7. In generating Design 3, we assumed as prior parameter estimates, the values discussed above. In Design 4 we double the magnitude of the prior parameter estimates, whereas Design 5 halves the magnitudes.

The sixth design was constructed also using the D-efficient criteria, however, a number of restrictions were placed on the design. Specifically, the design was generated such that the attribute levels for one of the non status-quo alternatives are always lower than that of the other non status-quo alternative. Given that higher levels for the non-price attributes are assumed to be more preferred (i.e., the prior parameters assumed were all positive for these attributes) whilst higher values for the price attribute are more preferred (i.e., a negative prior parameter) this design forces respondents to trade (simultaneously) the non-price attributes with price within each choice set. Such a constraint is designed to ensure that some form of trading always takes place in choice tasks. However, we note that strictly speaking one cannot assume that generating a design in this manner will avoid dominance in terms of preferences[1].

[1] Dominance implies that all respondents acting rationally will always select one alternative over all others present. Design 6 ensures that respondents will be faced with a comparison between a lower '*quality*' lower price alternative and a higher *quality* higher price alternative, but says nothing about the probability that one of the alternatives will be chosen. To establish whether an alternative is dominated or not, the analyst would need to calculate the choice probabilities (which are function of the design attributes and (prior) parameters). Once the choice probabilities are determined, the analyst would need to establish some rule as to what constitutes a dominated alternative based on the expected choice probabilities (e.g., if the probability is less than 0.1).

Designs 7 and 8 are orthogonal fractional factorial designs. In constructing the designs, no orthogonal design could be found that allowed for zero correlations both within and between the attributes of alternatives. As such a sequential design process was employed (see Louviere et al. 2000). This process involves first constructing an orthogonal design for alternative 1, and then using the same design to construct alternative 2. The process ensures that the designs are orthogonal in the attributes within alternatives, but not between alternatives. Given that the experiment is assumed to be unlabeled, the between alternative correlations are not of concern and hence the design process is appropriate. Whilst maintaining the (within alternative) orthogonality constraint, the D-efficient criteria was also applied to Design 8.

Designs 9, 10 and 11 are non-orthogonal designs generated to minimise respectively A- (Eq.(6), S- (Eq.(10) and B- (Eq.(7)) efficiency criteria. The remaining designs are generated in such a way as to minimise the sum of the C-efficiency measures (Eq.(20)). Designs 12 and 14 consider only the variances of the WTP values for the design attributes, whereas Designs 13 and 15 also consider the variance of the WTP for the status-quo constant. To illustrate the flexibility afforded by applying the C-criterion we use different weights for the variances of the WTPs of different attributes, when generating the last two designs, so that the criterion employed is the minimization of the weighted sum of the variance components of the attribute WTPs. This flexibility may be important in practice when the object of a stated preference study is to specifically calculate the WTP for a subset of the design attributes. The full set may include attributes considered important within respondent's preference space, but irrelevant from the viewpoint of WTP estimation. Alternatively, the absolute magnitudes of the WTP outcomes may also guide whether weighting should be applied, for example whether it is to be expressed in dollars or cents. For the present study, in constructing Design 14, attribute 1 is assigned the largest value of 0.4 because it is the one with lowest absolute WTP. As such, more precision (efficiency) is needed for this attribute relatively to the others to obtain a WTP estimate different from zero. For similar reasons attribute 2 is assigned a value of 0.35, and Attribute 3 of 0.25. The status-quo constant is ignored in this design, and hence has a weight of zero. A similar weighting procedure is applied in generating Design 15, with weights of 0.4, 0.3, 0.2 and 0.1 being applied to each of the design attributes and status-quo constant respectively.

8.3 Design Outcomes

Tables 2 and 3 present various efficiency measures for each of the 15 designs we generated. For each efficiency measure, excluding the B-efficiency measure, values are presented based on whether the constant term is considered in their calculation or not. As would be expected, the two random designs perform very poorly on each efficiency measure presented in the table. This outcome, however, is based on random chance, and different results might have been obtained if a different random allocation of the attribute levels were considered. The D-error design (Design 3) appears to perform very well on all criteria except B-error. According to the S-error for the design, a minimum of seven replications of the design (representing 140 choice observations) are required for all parameters, including the status-quo constant to be statistically significant at the 1.96 level. Of course this, number assumes that the prior parameter used is correct, hence, this represents only the theoretical minimum number of design replications that should be collected.

Table 2: Efficiency level outcomes for Designs 1 to 15

		D-error		A-error		C-error		Weighted C-error		S-error		B-error
Design	Effect	Without Constant	With Constant	Without Constant	With Constant	Without Constant	With Constant	Without Constant	With Constant	Without Constant	With Constant	
1	Base Design (random - unbalanced)	0.998	2.136	17.170	2.306	111.894	10.076	-	-	294.379	10.076	0.06%
2	Base Design (random - balanced)	0.920	1.398	4.847	2.202	22.677	8.630	-	-	59.270	9.218	1.67%
3	D-error	0.120	0.189	1.052	0.909	2.030	0.519	-	-	6.238	1.001	10.03%
4	Scale up (β×2)	0.198	0.290	3.612	3.895	0.656	0.176	-	-	9.529	1.015	7.77%
5	Scale down (β×0.5)	0.076	0.126	0.448	0.200	7.955	2.034	-	-	22.070	1.396	21.96%
6	Constrained trade-off	2.768	2.436	7.586	8.629	35.690	29.682	-	-	13.104	10.896	3.06%
7	Random orthogonal	1.847	1.640	5.282	5.602	20.037	12.398	-	-	15.368	8.786	16.21%
8	Efficient orthogonal	0.580	0.666	1.457	1.024	17.423	11.043	-	-	12.255	4.337	11.97%
9	A-error	0.212	0.283	0.653	0.526	2.503	1.230	-	-	4.456	0.943	10.25%
10	S-eff	0.330	0.369	1.218	1.353	2.471	1.327	-	-	2.596	1.782	18.78%
11	B-error	0.430	0.455	1.818	1.666	4.505	2.108	-	-	9.309	2.175	44.45%
12	C-error (attributes only)	0.153	0.281	4.282	2.984	6.456	0.455	-	-	36.386	3.293	7.53%
13	C-error (attributes + SQ)	0.206	0.262	2.838	3.185	1.454	0.551	-	-	5.585	3.454	21.42%
14	Weighted C-error (attributes only)	0.244	0.302	5.120	5.821	1.496	5.987	0.666	0.666	8.902	6.347	28.16%
15	Weighted C-error (attributes + SQ)	0.183	0.251	2.778	3.043	1.601	0.501	0.966	0.526	6.602	3.527	17.13%

Designs 4 and 5 represent the impact of assuming different prior parameter scales when generating the design. Contradictory results are produced when doubling and halving the magnitudes of the parameter priors. Halving the priors produce superior D- and A-error results, but dramatically worse results in terms of WTP and sample size requirements when compared to a doubling of the parameter prior scale. These results are counter-intuitive, as one would expect that doubling the assumed scale of the error (Design 4) and hence increasing the precision should lead to a higher efficiency. Instead, one observes the opposite. Increasing scale decreases the information content of the design for β while it increases it for the attribute WTPs. One possible cause for this might be that in generating the design, the attribute levels used are the same as those used for Design 3, and since it is differences in utility that matter most, the utility differences observed with a scaled up set of β are larger and induce large variations in choice probabilities, at the expenses of design balancedness and information content. The opposite may also explain findings for Design 5 where the scale of the priors is half that of Design 3.

Ignoring the randomly generated designs and designs where the parameter priors have been re-scaled, Design 6 performs quite poorly based on all criteria when compared to the other designs. This is because the trade-off constraint, whilst attempting to conform to some analyst imposed behavioural heuristic, fails to consider the statistical requirements that improve the statistical efficiency of experimental designs. In particular, the AVC matrix of a design, from which all efficiency measures are derived (save for the B-error measure), is the inverse of the second derivatives of the log-likelihood function for the design. As such, the AVC matrix is intrinsically related to the choice probabilities that the design will likely produce (given prior parameter estimates). In setting up the (behavioural) constraint, the expected choice probabilities for the design are also constrained, which in turn impacts on the design AVC matrix and its efficiency. As such, this design strategy, whilst behaviourally attractive, is likely to produce poor outcomes in terms of model results.

Design 7 represents the currently predominant method used for generating stated choice experimental designs. However, as shown here, the use of orthogonal designs tends to produce less than optimal outcomes in terms of expected model results, requiring larger sample sizes to retrieve statistically significant parameter estimates than other non-orthogonal designs. Design 8 represents an improvement on Design 7 by employing an algorithm that minimises the D-error of the design whilst maintaining orthogonality. Even so, the imposition of orthogonality represents a constraint on the efficiency of stated choice designs, for the exact same reasons as given for Design 6 poor performance. That is, the imposition of orthogonality only relates to the correlation structure of the design, but says nothing of the choice probabilities and hence AVC matrix that the design will likely produce[2].

Designs 9 to 11 were constructed so as to minimise A-, S and B- errors respectively. In each case, the designs produce the lowest (highest for the B-error design) values for the criteria for which the design was optimised. These designs appear to perform very similarly on all other criteria, however, the B-error design (Design 11) appears to require a larger minimum number of design replications in order to retrieve statistically significant parameter and WTP values. This finding is consistent with Sandor and Wedel (2001, 2002, 2005) and Kanninen (2002) who demonstrated that complete utility balance, as explored by Huber and Zwerina (1996), will result in sub-optimal designs.

[2] This statement is strictly not true. An orthogonal design will be optimal when all parameter priors are assumed to be zero (that is not important in the decision process). As such, orthogonal designs will only require the smallest possible design replications relative to all other designs when one is willing to assume that the attributes in the design do not play a role in the observed choices.

Table 3: T-ratio (assuming a single design replication) and minimum design replication requirements by attribute for Designs 1 to 15

	β		WTP		β		WTP		β		WTP	
	|t-*values*|	n	|t-*values*|	n	|t-*values*|	n	|t-*values*|	n	|t-*values*|	n	|t-*values*|	n
	Design 1	*Random allocation (unbalanced)*			*Design 2*	*Random allocation (balanced)*				*Design 3*	*D-efficient*	
Constant	0.255	59.270	0.267	53.962	0.114	294.379	0.106	340.096	0.785	6.238	0.814	5.804
β1	0.646	9.218	0.858	5.219	1.113	3.103	0.737	7.068	1.959	1.001	3.589	0.298
β2	1.427	1.886	1.347	2.118	1.681	1.359	0.647	9.173	2.000	0.960	5.016	0.153
β3	1.712	1.311	1.333	2.163	1.434	1.868	0.867	5.111	2.061	0.904	5.642	0.121
β4	0.854	5.268	n.a.	n.a.	0.617	10.076	n.a.	n.a.	1.977	0.983	n.a.	n.a.
	Design 4 $\beta\times 2$				*Design 5* $\beta\times 0.5$				*Design 6 Trade-off constrained*			
Constant	0.635	9.529	0.721	7.385	0.417	22.070	0.411	22.749	0.541	13.104	0.408	23.079
β1	1.951	1.010	7.294	0.072	1.659	1.396	1.744	1.263	0.612	10.243	0.567	11.961
β2	1.961	0.999	8.223	0.057	2.194	0.798	2.649	0.548	0.594	10.896	0.578	11.517
β3	1.966	0.994	9.596	0.042	2.279	0.740	2.816	0.484	0.689	8.097	0.786	6.222
β4	1.945	1.015	n.a.	n.a.	1.893	1.072	n.a.	n.a.	0.807	5.904	n.a.	n.a.
	Design 7 Orthogonal				*Design 8 Orthogonal efficient*				*Design 9 A-efficient*			
Constant	0.500	15.368	0.362	29.347	0.560	12.255	0.396	24.509	0.928	4.456	0.886	4.891
β1	0.753	6.769	0.890	4.847	1.732	1.281	1.137	2.970	2.018	0.943	2.186	0.804
β2	0.964	4.130	1.079	3.302	2.087	0.882	1.142	2.944	2.632	0.554	3.171	0.382
β3	0.800	6.001	1.081	3.286	2.290	0.733	1.118	3.075	2.868	0.467	3.802	0.266
β4	0.661	8.786	n.a.	n.a.	0.941	4.337	n.a.	n.a.	2.310	0.720	n.a.	n.a.
	Design 10 S-efficient				*Design 11 B*-efficient				*Design 12* C_p-efficient attr. only			
Constant	1.217	2.596	0.935	4.393	0.642	9.309	0.646	9.210	0.325	36.386	0.408	23.051
β1	1.495	1.718	1.891	1.074	1.329	2.175	1.675	1.369	1.080	3.293	3.881	0.255
β2	1.693	1.340	3.021	0.421	1.584	1.532	2.344	0.699	1.109	3.122	5.323	0.136
β3	1.702	1.326	3.843	0.260	1.501	1.705	2.966	0.437	1.133	2.995	6.025	0.106
β4	1.468	1.782	n.a.	n.a.	1.400	1.959	n.a.	n.a.	1.109	3.125	n.a.	n.a.
	Design 13 C_p-efficient attr. + sq				*Design 14 Weighted* C_p-efficient attr. only				*Design 15 Weighted* C_p-efficient attr. + sq			
Constant	0.829	5.585	1.052	3.470	0.657	8.902	1.023	3.671	0.763	6.602	0.954	4.223
β1	1.064	3.391	3.333	0.346	0.778	6.347	3.478	0.318	1.044	3.527	3.525	0.309
β2	1.055	3.454	4.714	0.173	0.800	5.995	4.563	0.185	1.105	3.145	5.127	0.146
β3	1.100	3.175	5.662	0.120	0.807	5.904	5.826	0.113	1.122	3.051	5.786	0.115
β4	1.103	3.156	n.a.	n.a.	0.801	5.987	n.a.	n.a.	1.097	3.192	n.a.	n.a.

Our last group of comparisons are made across designs obtained by using various specifications of C-efficiency as the optimization criteria. These designs perform well compared to most other designs, however, a number of issues arise which require further discussion. Firstly, the theoretical minimum number of design replications required for Design 12 is 37 (740 choice observations) if all parameters are to be found to be statistically significant as per Eq.(10) (assuming the priors have been correctly specified). Table 3, demonstrates the asymptotic *t*-ratios for each attribute and WTP for each design, as well as the number of design replications required in order for the asymptotic *t* -ratios to be greater than 1.96. An examination of this Table for Design 12 shows that the requirement for 37 replications of the design is a result of the status-quo constant, which was not considered when generating the design. As such, it is questionable as to whether one would consider 37 replications or the next highest value of four replications to be the minimum.

A second observation relates to the use of the C-efficiency criteria as expressed previously. The C-efficiency criteria as implemented here relates only to the variances of the ratios of two parameters, and not the variances of the parameters themselves. Whilst there exists a relationship between the two, the additional non-variance terms contained within Eq.(20) may compensate for larger parameter variances when minimising the equation. As such, it may be possible to minimise the variance of the ratio of the two parameters whilst obtaining a relatively large variance for one or more of the parameters themselves. This has implications when calculating the WTP for that attribute and it is clearly demonstrated in Table 3. Consider for example, Design 13. For the status-quo constant term to achieve an asymptotic *t*-ratio of 1.96, at least six (rounding up from 5.585) design replications are required (120 choice observations), whereas only four (rounding up from 3.470) replications are required (80 choice observations) for the WTP for the status-quo constant term to achieve statistical significance. Given that the WTP for an attribute should only be calculated if the individual parameters are statistically significant, the higher value of the two should be used (i.e., six design replications). A search through Table 3 reveals that Designs 9 (A-efficiency) and 10 (S-efficiency), whilst requiring a larger number of design replications for all WTP values to become statistically significant, would require only five design replications (i.e., 100 choice observations) for all parameter and WTP values to be statistically efficient. As such, these designs would be preferred based on these criteria.

Whilst we do not implement it here, it should be possible to create a new optimisation criterion similar to the S-efficiency measure that minimises the largest sample size required for the ratios of two parameters to be statistically significant. Indeed, one could combine this with the current S-efficiency measure, and jointly minimise both.

7. Conclusion and direction of further research

The use of stated preference methods has become increasingly accepted in the policy arena as a way to investigate non-market values worldwide. Yet, choice modelling has not been subject to the degree of investigation and scrutiny dedicated to contingent valuation in the nonmarket valuation literature. With particular regards to the topic of experimental design tailored to the specific needs of non-market valuation practitioners the literature is still scarce. This study had the objective of bringing together a number of considerations and statistics that the practitioner could find of

interest. In particular, the principles outlined here can be adopted in the evaluation of choice model designs predicated under different assumptions from the one used for convenience here as the main example.

C-efficiency, for example, is a criterion for design evaluation that although proposed over 15 years ago, is still rarely used. Sample size determination, as we explained here can be theoretically linked to design properties, and can itself be used as a criterion for design search. Importantly, we suggest alternative ways of reporting design statistics in applied studies that go beyond the frequently used percent efficiency criterion originally proposed for multivariate linear regression studies explaining treatment effects in agricultural experiments. We show how this criterion is irrelevant and a bad proxy for C-efficiency, which ought to be what matters when the focus is WTP estimation.

We have intentionally neglected several important considerations related to the behavioural efficiency of the design, concentrating our focus on the statistical efficiency and the comparison of different criteria to practically measure it. Future research should focus on respondent efficiency as well. Although perhaps the current level knowledge on how respondents process the information provided in choice tasks is still insufficient to derive efficiency measures to evaluate behavioural efficiency, this knowledge gap is filling quickly. For example, extensive research has been conducted on the impact upon behavioural responses given various design dimensions. For example, the number of alternatives within the task (Hensher et al. 2001), the number of attributes (Pullman et al. 1999), the number of attributes and alternatives (Arentze et al. 2003; DeShazo and Fermo 2001), the impact of attribute level range upon response (Cooke and Mellers 1995; Ohler et al. 2000; Verlegh et al. 2002) and the number of choice profiles shown to respondents (Brazell and Louviere 1998) have all been examined. More recently, Hensher (2004, 2006a,b) and Caussade et al. (2005) examined all of the above effects simultaneously. Nevertheless, an examination of the combination of the design and respondent efficiency remains to date, ever elusive.

References

Alberini, A. Optimal Designs for Discrete Choice Contingent Valuation Surveys: Single-Bound, Double Bound and Bivariate Models. *Journal of Environmental Economics and Management*, 1995, 28, 287-306

Arentze, T., Borgers, A., Timmermans, H., and DelMistro, R. (2003) Transport stated choice responses: effects of task complexity, presentation format and literacy, *Transportation Research Part E,* 39, 229–244.

Bliemer, M.C.J. and Rose, J.M. (2005) Efficiency and Sample Size Requirements for Stated Choice Studies, working paper: ITLS-WP-05-08.

Bliemer, M.C.J., and Rose, J.M. (2006) Designing Stated Choice Experiments: State-of-the-art, paper presented at the 11th International Conference on Travel Behaviour Research, Kyoto, Japan.

Bliemer, M.C.J., Rose, J.M. and Hensher, D.A. (2007) Constructing Efficient Stated Choice Experiments Allowing for Differences in Error Variances Across Subsets of Alternatives, accepted for *Transportation Research B.*

Brazell, J.D. and Louviere, J.J. (1998) Length effects in conjoint choice experiments and surveys: an explanation based on cumulative cognitive burden. Department of Marketing, The University of Sydney, July.

Breffle, W. S. & Rowe, R.D. (2002) Comparing Choice Question Formats for Evaluating Natural Resource Tradeoffs, *Land Economics*, 78, 298-314

Burgess, L. and Street, D.J. (2005) Optimal designs for choice experiments with asymmetric attributes, *Journal of Statistical Planning and Inference*, 134, 288-301.

Caussade, S., Ortúzar, J. de D., Rizzi, L.I. and Hensher, D.A. (2005) Assessing the influence of design dimensions on stated choice experiment estimates, *Transportation Research B*, 39, 621-640.

Cook, R.D., and Nachtsheim, C.J. (1980) A comparison of algorithms for constructing exact *D*-optimal designs. *Techometrics* 22, 315-324.

Cooke, A.D. and Mellers, B.A. (1995) Attribute Range and Response Range: Limits of Compatibility in Multiattribute Judgment, *Organizational Behavior and Human Decision Processes*, 63(2), 187-194.

DeShazo, J.R. and Fermo G. (2002) Designing choice sets for stated preference methods: the effects of complexity on choice consistency, *J Environmental Economics and Management*, 44(1), 123-143.

Ferrini, S. and Scarpa, R. (2007) Designs with a-priori information for nonmarket valuation with choice-experiments: a Monte Carlo study, *Journal of Environmental Economics and Management*, 53, 342-363.

Hensher, D.A. (2006a) Revealing differences in behavioural response due to the dimensionality of stated choice designs: an initial assessment. *Environmental and Resource Economics*, 34, 7-44.

Hensher, D.A. (2006b) How do respondents process stated choice experiments? attribute consideration under varying information load. *Journal of Applied Econometrics*, 21, 861-878.

Hensher, D.A. (2004) Accounting for stated choice design dimensionality in willingness to pay for travel time savings. *Journal of Transport Economics and Policy*, 38, 425-446.

Hensher, D.A., Stopher, P.R. and Louviere, J.J., (2001) An Exploratory Analysis of the Effect of Numbers of Choice Sets in Designed Choice Experiments: An Airline Choice Application, *Journal of Air Transport Management*, 7, 373-379.

Huber, J. and Zwerina, K. (1996) The Importance of Utility Balance in Efficient Choice Designs, *Journal of Marketing Research*, 33, 307-317.

Kanninen, B. J. (1993a) Optimal Experimental Design for Double Bounded Dichotomous Choice Contingent Valuation *Land Economics,* 69, 138-146

Kanninen, B. J. (1993b) Design of sequential experiments for CV studies *Journal of Environmental Economics and Management*, 1993b, 25, 1-11.

Kanninen, B. J. (2002) Optimal Designs for Multinomial Choice Experiment *Journal of Marketing Research*, 39, 214-227

Kessels, R., P. Goos, and x Vandebroek, R. (2006) A comparison of criteria to design efficient choice experiments, *Journal of Marketing Research*, 43(3), 409-419.

Louviere, J.J., Hensher, D.A. and Swait, J.D. (2000) *Stated Choice Methods—Analysis and Application*. Cambridge University Press, UK.

Louviere, J.J., Street, D.J. and Burgess, L. (2003), A 20+ Years Retrospective on Choice Experiments, Ch8 in Marketing Research and Modeling: Progress and Prospects," Wind, Y. and Green, P.E. (eds), New York: Kluwer Academic Press, 201-214.

Lusk, J.L. and Norwood, F.B. (2005) effect of experimental design on choice-based conjoint valuation estimates, *American Journal of Agricultural Economics*, 87(3), 771–785.

Ohler, T., Li. A., Louviere, J.J., and Swait, J. (2000) Attribute range effects in binary response tasks *Marketing Letters*, 11, 3, 249-260.

Pullman, M.E., Dodson, K.J., and Moore, W.L., (1999) A Comparison of Conjoint Methods When There Are Many Attributes, *Marketing Letters*, 10(2), 1-14.

McFadden, D. (1974) Conditional logit analysis of qualitative choice behaviour, In Zarembka, P. (ed.), *Frontiers in Econometrics*, Academic Press, New York, 105-142.

Sándor, Z., and Wedel, M. (2001) Designing Conjoint Choice Experiments Using Managers' Prior Beliefs. *Journal of Marketing Research* 38, 430-444.

Sándor, Z., and Wedel, M. (2002) Profile Construction in Experimental Choice Designs for Mixed Logit Models, *Marketing Science* 21(4), 455-475.

Sándor, Z., and Wedel, M. (2005) Heterogeneous conjoint choice designs. *Journal of Marketing Research* 42, 210-218.

Scarpa, R.; Ferrini, S. and Willis, K.G. (2005) Performance of error component models for status-quo effects in choice experiments, Ch13 in *Applications of simulation methods in environmental and resource economics,* Scarpa, R. and Alberini, A. (eds), Springer, 247-274.

Scarpa, R., Campbell, D. and Hutchinson, W.G. (2007) Benefit estimates for landscape improvements: Sequential Bayesian Design and Respondent's Rationality in a choice experiment, *Land Economics*, 83(4), 617-634.

Street, D.J., and Burgess, L. (2005) Optimal and near-optimal pairs for the estimation of effects in 2-level choice experiments, *Journal of Statistical Planning and Inference*, 118, 185-199.

Street, D.J., Bunch, D.S. and Moore, B.J. (2001) Optimal designs for 2^k paired comparison experiments, *Communications in Statistics, Theory, and Methods*, 30(10), 2149-2171.

Street, D.J., Burgess, L. and Louviere, J.J. (2005) Quick and easy choice sets: Constructing optimal and nearly optimal stated choice experiments, *International Journal of Research in Marketing*, 22, 459-470.

Train, K. (2003) *Discrete Choice Methods with Simulation*, Cambridge University Press, UK.

Verlegh, P.W., Schifferstein, H.N., and Wittink D.R. (2002) Range and Number-of-Levels Effects in Derived and Stated Measures of Attribute Importance, *Marketing Letters*, 13, 1, 41-52.