How good is good? Exploring frontiers of experimental design efficiency

Tom. Longworth
MWT Transport Planning, Sydney, NSW, Australia

1 Introduction

A fundamental input to transport planning is the representation of different economic agents' behaviour. Stated choice methods have become an accepted and widely used standard when attempting to represent behaviour of these agents. One of the more exciting recent developments in this field is a re-think of the basis of experimental designs, with the previous method of orthogonal designs losing favour to a range of new approaches. These new methods offer the opportunity to reduce sample sizes, whilst supporting the estimation of statistically robust models. A drawback of these approaches to experimental design is that it is difficult to know how close a particular design is to the best achievable design(s) for the specific experiment.

This paper reports a study that set out to explore how efficient a design was by generating a large set of random designs to see how efficient the best design was, and to provide a frequency distribution of a widely used measure of efficiency. Against this pool of random designs, a comparison was made with the designs generated using a genetic algorithm. The initial phase of the study reported in this paper was restricted to exploring multi-nominal logit (MNL) models with fixed prior estimates of model parameters; subsequent work will explore mixed logit models and error components models; it will also use Bayesian distributions as the prior estimates of model parameters. While this restricted scope substantially limits the applicability of the results, it does provide an indication of potential gains in design efficiency that might be achieved through the application of genetic algorithms to different types of stated choice experiment design and approaches to estimating prior values of model parameters; as well as establishing and proving procedures for computing and data handling for subsequent work.

2 Background

In stated choice analysis a survey containing a number of choice sets is presented to respondents. A number of choice sets are presented in which attributes of the various alternatives are varied to measure respondent preferences. Choices made by the respondents are then recorded and analysed, permitting the estimation of models to represent the choice behaviour observed by the survey. This process is described in a wide body of literature, including Hensher, Rose and Greene (2005), Louviere, Hensher and Swait (2000) and Train (2003).

The experimental design process derives the particular choice sets to be administered to survey respondents. The number of alternatives, attributes and their levels, as well as the number of choice sets are determined as part of the design process, and their determination is based on requirements of the analytical framework and the analysts' experience. Once these aspects of the design are determined, the next step is to select the specific choice sets that are to be administered in the experiment. The selected choice sets comprise the experimental design; this is but one design amongst a very large number of potential designs.

The process that was commonly adopted until the last five or so years was the use of orthogonal designs, based on the experimental design literature. One method was to consult design catalogues, such as Hahn and Shapiro (1966), to extract the appropriate design. The

designs in the catalogues provide statistically efficient designs, avoiding cross-correlations in the experimental results.

However, many of the experiments for which these cookbook designs were developed had outputs that showed linear or near-linear responses to input stimuli. In stated choice analysis outputs are related to inputs through a highly non-linear model, as shown in Hensher, Rose and Greene (2005).

$$\text{Prob}_i = \frac{\exp V_i}{\sum_{j=1}^{J} \exp V_j}$$

The literature notes that the result of this non-linearity in the model is that orthogonality in the design may not translate to orthogonality in the data. Consequently a number of different approaches have been developed. These approaches include, among others, optimal designs which are described by Street, Burgess and Louviere (2005), efficient designs which are described by Rose and Bliemer (2006), and utility balanced designs which are described by Huber and Zwerina (1996). In these methods, knowledge about the analytical process is applied to the selection of the individual choice sets to be administered in the experiment.

Other papers in the area of experiment design for stated choice methods include, Bliemer, Rose and Hensher (2007); Carlsson and Martinsson (2002); Ferrini and Scarpa (2007); Kanninen (2002); Sandor and Wedel (2001); Sandor and Wedel (2002); Sandor and Wedel (2005); and Bliemer, Rose and Hess (2007).

Efficient designs use fore knowledge of the likely value of the parameters to be estimated by the analysis of the experimental results to search for a design in which the choice sets produce an asymptotic variance-covariance matrix (AVC) with minimum 'D-error' (matrix determinant raised to the power of the inverse of the number of parameters). Rose and Bliemer (2006) provide a detailed discussion of different measures of efficiency and their implications for choice analysis strategies. The D-error does not necessarily identify the design requiring the minimum sample size to maintain significant t-statistics, as the D-efficiency measure is based on the determinant of the AVC matrix. Whereas the S-efficiency measure, which is calculated from the AVC diagonal, does yield a design with the minimum sample size design (from a statistical perspective).

A potential problem with this approach is that prior knowledge of parameter values might be poor. For the purposes of this paper, it is assumed that prior knowledge of parameter values is good, although a number of strategies are available to deal with this and are described in Rose and Bliemer (2006). These strategies include conduct of small pilot surveys to hone the prior values, as well as use of Bayesian distributions as prior estimates of model parameters, which are sampled from a distribution of potential parameter values.

Another potential problem with this approach is that the analyst does not know how good their selected design is, compared with other, as yet unidentified, designs. The very large set of possible permutations of choice sets which could be used as experimental designs makes a comprehensive exploration of every design impractical for all but the simplest choice situations. By very large, it is meant that, even for small experiments, with existing computers, it is not considered practicable to comprehensively evaluate all designs. Some computing time estimates are reported in the following sections, in support of this claimed intractability.

A number of different search methods are discussed in the literature, including relabelling, swapping and cycling (RSC) described in Huber and Zwerina (1996) and genetic algorithms described in Hensher, Rose and Bliemer (2007). Another method is to randomly generate designs and keep the best (i.e., minimum D-error design, if D-efficiency is the design

criteria). One approach to this method is described in detail by Hensher, Rose and Bliemer (2007), in which analysis is implemented in a spreadsheet package in a manner that generates designs with attribute balance, and the design with the best D-error is retained, until it too is surpassed and replaced by a better design.

This paper reports the findings of a study that, in its first stage, took two experimental designs and drew at random a large number of experimental designs for each. While not a comprehensive set of experimental designs for the two experiments, it does provide a large pool of designs and an indication of the level of effort required to generate reasonable D-error designs. Stage two of the study compares the random designs from stage one with designs derived using a genetic algorithm. From this it was possible to:

- Find a number of efficient designs for each experiment
- By randomly sampling across the range of possible designs, a frequency distribution for the measure of efficiency (D-error) was prepared for each experiment
- Explore how the application of genetic algorithms might improve selection of D-efficient designs by comparison of results with randomly drawn designs

3 Method – Stage 1 Random Designs

Two separate choice experiments were explored. The first is a three alternative experiment which is described in Hensher, Rose and Bliemer (2007). This design was chosen as a benchmark design. The specification for the design is repeated below, and it is referred to subsequently as Experiment 1:

$$U_1 = \beta_{10} + \beta_1 x_{11} + \beta_2 x_{12} + \beta_{13} x_{13}$$
$$U_2 = \beta_{20} + \beta_1 x_{21} + \beta_2 x_{22} + \beta_{23} x_{23}$$
$$U_3 = \beta_1 x_{31} + \beta_{33} x_{33}$$

With attribute levels of:

$$x_{11}, x_{21}, x_{31} \in \{6,8,10,12\},$$
$$x_{12}, x_{22} \in \{4,8\},$$
$$x_{13}, x_{23}, x_{33} \in \{0,1\}$$

Eight choice situations are assumed. Prior estimates of parameters are:

$$U_1 = 1.2 - 0.6x_{11} - 0.4x_{12} + 0.3x_{13}$$
$$U_2 = 0.8 - 0.6x_{21} - 0.4x_{22} + 0.8x_{23}$$
$$U_3 = -0.6x_{31} - 1.0x_{33}$$

A second experiment was formulated for this study, this time with two alternatives. It is referred to subsequently as Experiment 2 and is described by:

$$U_1 = \beta_{10} + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13}$$
$$U_2 = \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{23}$$

With attribute levels of:

$$x_{11}, x_{21} \in \{15,20,25,30\},$$
$$x_{12}, x_{22} \in \{4,8\},$$
$$x_{13}, x_{23} \in \{4,8\}$$

Again, eight choice situations are assumed. Prior estimates of parameters are:

$$U_1 = -3.5 - 0.95x_{11} - 2.3x_{12} - 1.8x_{13}$$
$$U_2 = -0.95x_{21} - 2.3x_{22} - 1.8x_{23}$$

While Experiment 2 is simpler and, hence, requires less computational effort than the first design, it is still regarded as meaningful. In a transport context this experiment could represent a choice between two different surface public transport alternatives for a mid-length trip, with the first attribute representing travel time, the second attribute representing an access measure and the third attribute representing an egress measure. Conceivably this experiment might represent choices made by people who find their trip ends in between and in betwixt two strong near parallel bus or tram corridors, and are faced with the trade-off of a longer access/egress combination for a shorter travel time, or vice versa.

The total number of possible designs for each of the experiments is:

- Experiment 1 – 2048 choice sets with 8 choice sets in an experiment: 7.57E+21, with attribute balance as a constraint, this reduces to around 2.4E+17
- Experiment 2 – 256 choice sets with 8 choice sets in an experiment: 4.10E+14, applying the constraint of attribute balance reduces this to about 2.5E+11

The analysis entailed the generation of a large number of designs for each of the experiments and the calculation of the D-error for each of the designs. This applied the analytical framework outlined in detail by Hensher, Rose and Bliemer (2007), although its random designs were drawn differently, and the application environment was different. In brief, the framework:

- Uses prior estimates of parameters (as assumed above) to calculate the utilities for each of the alternatives in each of the choice sets, this yields the probability of choosing each alternative in the choice set;
- These choice probabilities are then multiplied by the attribute levels;
- The products of this process are summed for each attribute for each choice set;
- These totals are deducted from the attribute value, multiplied by probabilities and then their square roots are taken;
- The resulting matrix is multiplied by its transpose to produce the Fisher Information matrix;
- This is inverted to produce the AVC matrix;
- The determinant of the AVC matrix is raised to the inverse of the number of parameters, yielding the D-error.

A software routine was written for this study for each of the two experiments. This software:

- Generated all choice sets for each of the experiments, assigning an index number to each choice set and holding these in memory.
- It then generated sets of eight unique random numbers for each index number of each choice set (i.e. between 1 and 2048 for Experiment 1 and between 1 and 256 for Experiment 2). The resulting experimental design was tested for attribute balance. If the design had attribute balance, then the experimental design was analysed, computing the D-error and storing the design's index numbers and D-error for post processing.
- Post processing entailed the consolidation of all output files and their subsequent interrogation to produce a frequency distribution of D-errors and to trap the most efficient designs. The cut-off for trapping such 'good' designs was determined after an initial exploration of the results.

Due to power system unreliability, the application was structured to write out a file of results approximately every hour to minimise loss of data, should a power interruption occur. As the

application was compiled as a standalone executable file, it could be run on multiple personal computers simultaneously, so long as the personal computer had a compatible operating system and sufficient disk space – no other applications were required on the machine. As each of the copies of the software was given its own seed for random number generation, there would be few individual designs generated and analysed more than once.

It should be noted that different experiments would not be expected to have the same D-error distributions. That is, the results are unique to a specific choice situation. Also different prior values of the parameters would be expected to produce different D-error distributions and minimum D-error values.

An advantage of this structure of processing was that it permitted unused computers to be put to work for short periods, contributing to the processing task. A disadvantage was that an overwhelming proportion of designs sampled at random had unbalanced attributes, so that a very large number of random draws were required to produce a small number of balanced designs. Despite this disadvantage, this sampling process was retained to permit possible longer term exploration of D-error variability as one moved systematically along the 'choice set axis'; by this it is meant that each experimental design can be allocated to an index comprising the individual choice set index numbers ([cset1_index], [cset2_index], [cset3_index],…,[cset8_index]) along the x-axis and plotted against the D-error. At this stage of the current study the 'choice set axis' is a rather long axis (for Experiment 1 it is approximately 3.09E+26 ordinals long) and has yet to be explored.

An extension of the method was to sample and analyse a small set of designs with unbalanced attributes. A choice experiment design will generally have attribute balance as a specified constraint, so that respondents deal with each attribute level the same number of times. Nonetheless it was of interest to see how efficient these designs might be.

4 Results – Stage 1 Random Designs

The table below summarises the frequency distribution of D-errors for the two experiments with and without attribute balance, and includes the best D-error computed in stage one by the random sampling approach. It also reports an approximate sample rate for each of the experiments.

Table 1 - Summary of Stage One results, by experiment and by attribute balance condition

D-error	Experiment 1				Experiment 2			
	Attribute balance		Attribute unbalance		Attribute balance		Attribute unbalance	
	#	%	#	%	#	%	#	%
<0	69,826	0.06%	16,496	0.25%	129,168	0.02%	37,365	0.01%
>=0 & <1	4,669,257	3.88%	117,721	1.80%	24,990,158	4.47%	13,555,884	4.45%
>=1	115,585,917	96.06%	6,390,580	97.94%	533,319,979	95.50%	291,138,774	95.54%
Total	120,325,000	100.00%	6,524,797	100.00%	558,439,305	100.00%	304,732,023	100.00%
Best D-error*	0.75087		0.78728		0.22201		0.22355	
Sample rate	1 in 1.96E+09		1 in 1.16E+15		1 in 4.5E+02		1 in 1.34E+06	

*Note * best D-error found to date; actual best D-error is not known.*

Clearly there are relatively few 'good' designs, with only between two and five percent of designs having a D-error within the broad area of interest. Of note is that the D-error distribution of the designs without attribute balance is broadly similar to those with attribute balance. Also the 'best' D-error is not vastly different between cases with and without attribute balance in each of the two experiments; possibly the difference is accounted for by the smaller sample size in the cases without attribute balance. This suggests that achieving an efficient design would not be a valid reason to maintain attribute balance, although the reason for attribute balance outlined above obviously does not change.

Experiment 2's higher proportion of experimental designs with D-errors between 0 and 1 reflects the lower D-error of good designs for that experiment, somewhere around 0.22 compared with 0.75 for Experiment 1.

Despite the apparently large sample sizes, the sample rates are small, especially for Experiment 1. In the course of this project the software would produce about 50,000 to 100,000 analysed balanced designs per hour for Experiment 1, depending on the specification of the machine used. A comprehensive analysis of Experiment 1's designs with attribute balance, assuming the approach could be improved to generate one million analysed designs per hour, would take over twenty million processor years. Even if it could run one thousand times faster still, the requirement for twenty thousand processor years is not an insubstantial level of resource. This is considered beyond current practical computing capability, and certainly beyond this study's budget.

To provide an appreciation of the shape of the frequency distribution near the designs with desirable D-errors, charts were produced showing a 'coarse' frequency distribution for each of Experiment 1 and Experiment 2 with and without attribute balance, for D-errors between 0 and 20 with a resolution of 0.01. These are shown below.

Figure 1 – Experiment 1 – Frequency distribution (coarse) with attribute balance

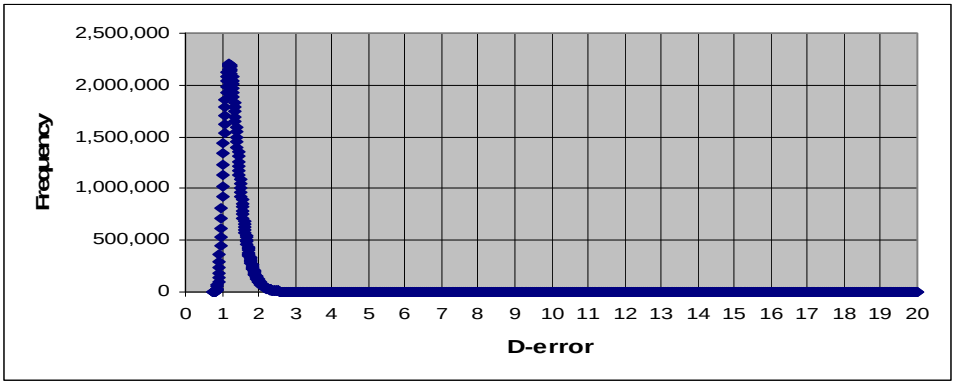


Figure 2 – Experiment 1 – Frequency distribution (coarse) without attribute balance

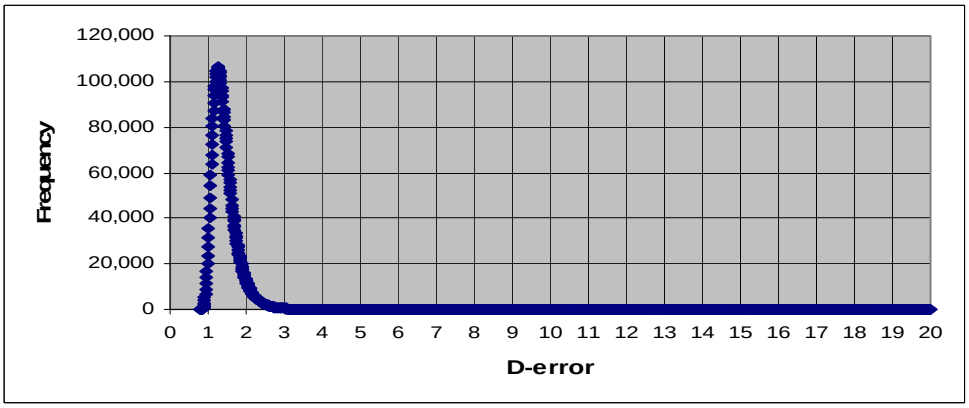


Figure 3 – Experiment 2 – Frequency distribution (coarse) with attribute balance

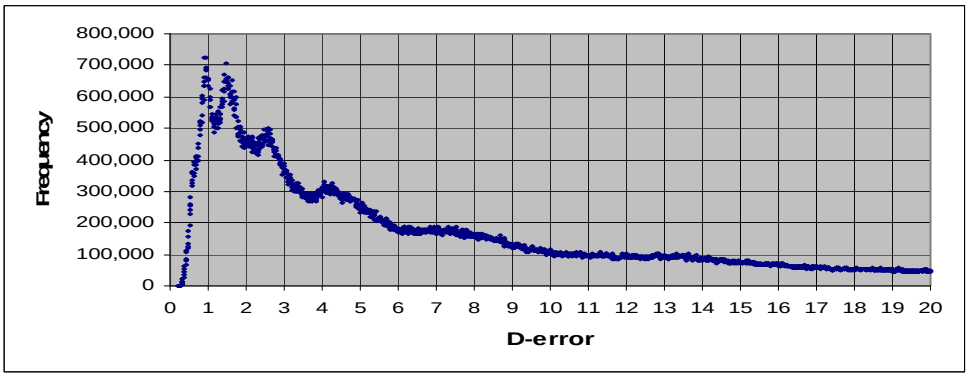
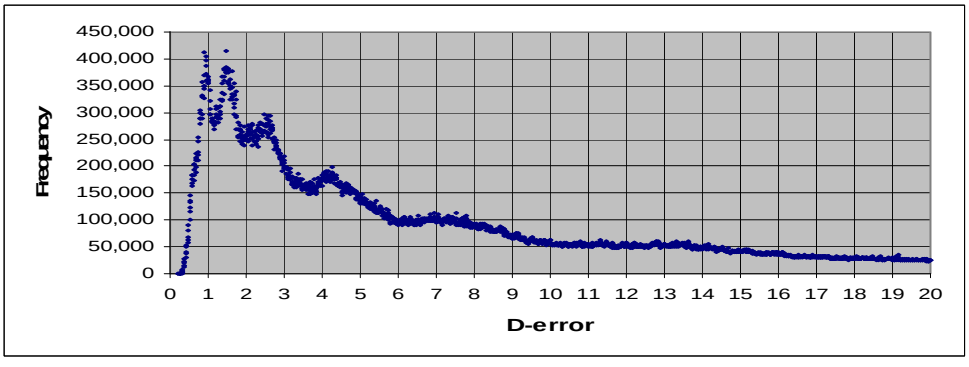


Figure 4 – Experiment 2 – Frequency distribution (coarse) without attribute balance



By inspection the above charts indicate:

- Broadly similar shaped distributions within each of the experiments with and without attribute balance.
- There are substantial differences in shape between experiments – Experiment 1 has few designs with D-errors greater than 4, and has a smooth surface; Experiment 2 has a high proportion of designs with D-errors out to 20 and beyond, it also has several localised peaks.
- The left hand end of all four charts suggest that the minimum D-error is probably not that far further to the left than identified.

In order to examine the frequency distributions close to the best D-errors found by the random method, charts were prepared at a D-error resolution of 0.001 for a range sufficient for the reader to identify the location of the best D-error found to date. As a consequence of these presentational considerations the scales of the charts are both horizontally and vertically inconsistent. Only designs with attribute balance are considered further. Note that D-error values for which no design was found are shown as points on the horizontal axis. These charts are shown below.

Figure 5 – Experiment 1 – Frequency distribution (fine) with attribute balance

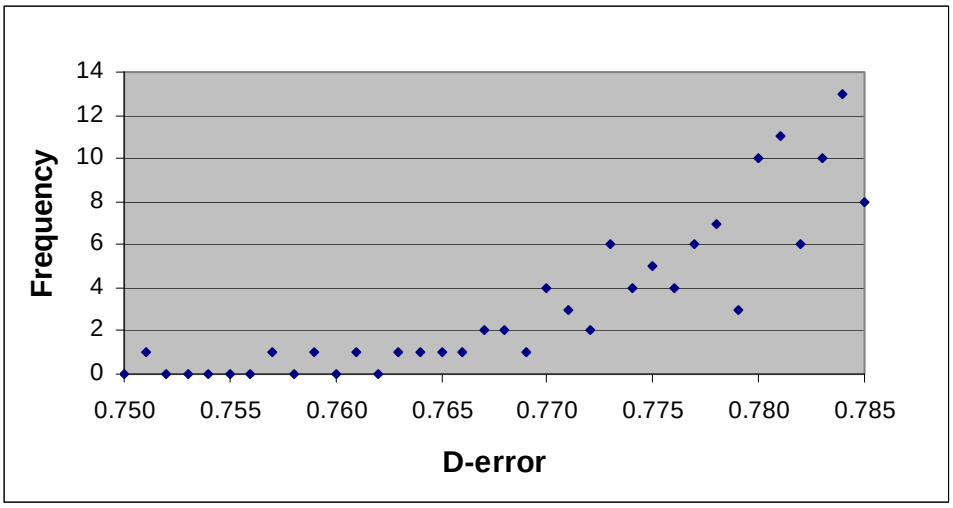
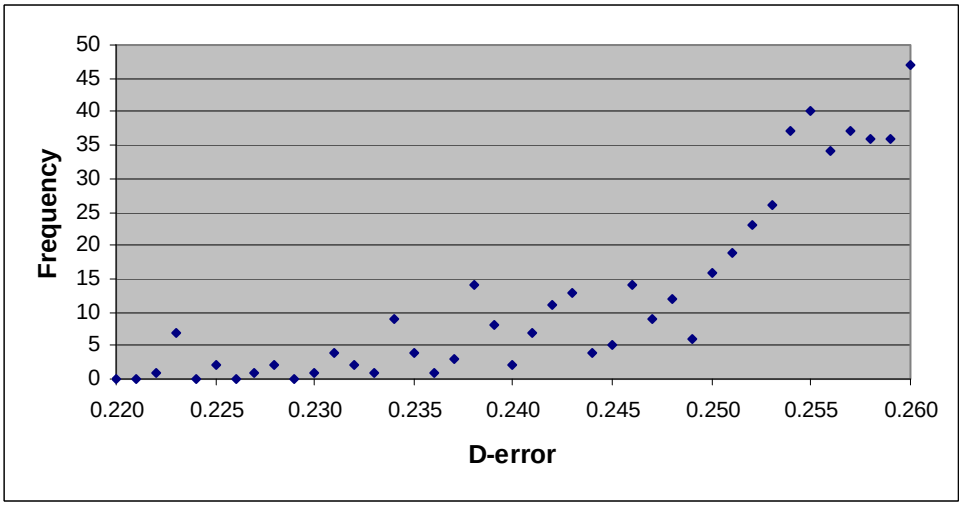


Figure 6 – Experiment 2 – Frequency distribution (fine) with attribute balance



Assuming that the samples drawn for both Experiment 1 and 2 (with attribute balance) are broadly representative of the comprehensive frequency distribution, it indicates that there are

potentially a considerable number of designs close to the 'best' designs found using this random method (and shown on Figures 5 and 6):

- Experiment 1 – D-error between 0.75 and 0.77 there were 13 designs in the sample, which might represent approximately 2.55E+10 potential designs in this part of the sample.
- Experiment 2 – D-error between 0.22 and 0.23 there were 14 designs, which might represent approximately 6.3E+03 potential designs in this part of the sample.

So, despite the very large samples drawn for both experiments, and the considerable computational effort involved, the random method has only scratched the surface in that part of the frequency distribution of designs that is of interest to the analyst. Furthermore, there may well be more designs below the minimum D-error generated by this method.

5 Method – Stage 2 Genetic Algorithm

The second stage of the investigation was to apply an alternative method to derive designs and compare the resultant D-errors and level of effort required with Stage One's method.

A routine was coded for each of the two experiments that implemented the following basic structure of design generation:

- Step 1 - 5,000 designs taken as the initial population.
- Step 2 - From the 2,500 pairs of designs, 8 progeny were bred for each pair. Designs were combined by randomly drawing attributes from the parent designs and combining to form an unaltered progeny design.
- Step 3 - Each progeny design was then altered by randomly choosing one or more of the attributes for alteration, and then swapping one of the attributes in a randomly selected pair of choice set/alternative combinations.
- Step 4 - The subsequent 20,000 progeny designs were evaluated and sorted. The 5,000 designs with the best D-errors were selected as the new generation from which to breed the next generation of designs.

This basic structure is described in Hensher, Rose and Bliemer (2007).

This process was seeded using 5,000 designs drawn at random, which were generated at in the first stage of the study. As there were also a large sample of 'good' designs already available, the best 5,000 designs from each experiment in stage one were also used as a starting point for comparison purposes.

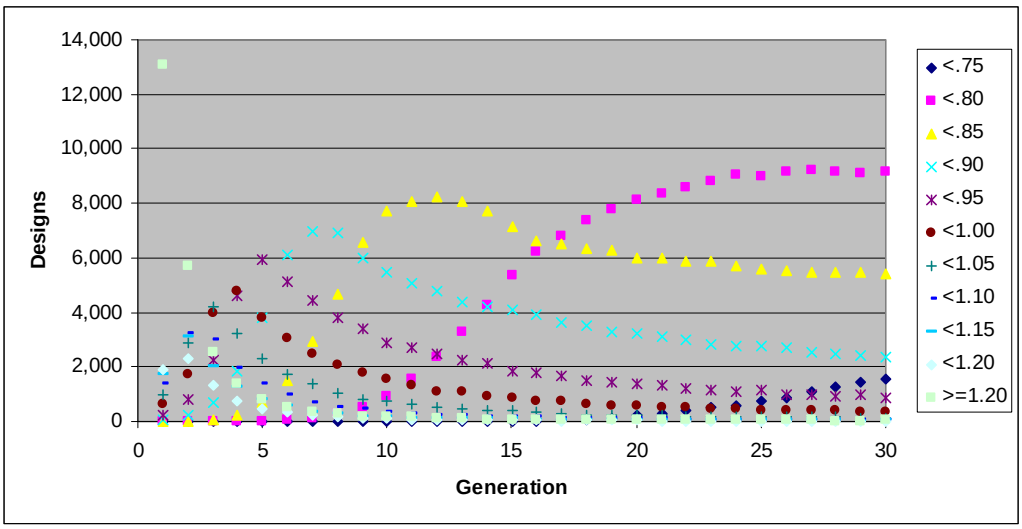
A total of 30 generations were developed and evaluated for Experiment 1.

6 Results – Stage 2 Genetic Algorithm

6.1 Experiment 1

The following chart shows the progressive frequency distribution of D-errors for each generation of new designs. This is based on using 5,000 randomly generated designs from Stage 1 of the study.

Figure 7 - Experiment 1 - Initial population randomly generated, comparison of D-error frequency bands by generation

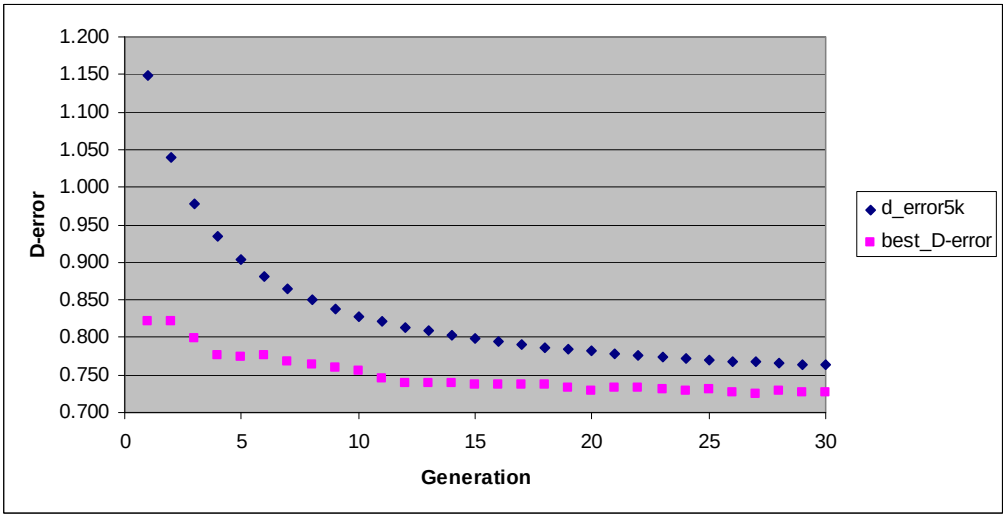


Key features of the chart are:

- The very rapid decay in the number of 'bad' designs with D-errors greater than 1.20, falling from just over two-thirds of designs in the first generation to less than 10% by generation 4.
- The large number (approximately 8,000) of designs produced by generation 20 with D-errors less than 0.80 (only 691 designs, with attribute balance, with D-errors less than 0.80 were generated in Stage 1 of the study, from a total of over 120 million designs) – by generation 20 only 400,000 designs have been evaluated and yet has produced more than ten times the number of 'good' designs.
- Whilst partly obscured by other data points, from about generation 24 the frequency of designs with D-errors less than 0.75 (i.e., better than the best D-error design from Stage 1) is becoming obvious with nearly 2,000 designs of this standard generated by generation 30.

The following chart compares the best D-error and the D-error value with 5,000 better designs (breeding cut off) for each generation.

Figure 8 – Experiment 1 – initial population randomly generated, progress of best D-error and D-error of 5000th ranked design by generation

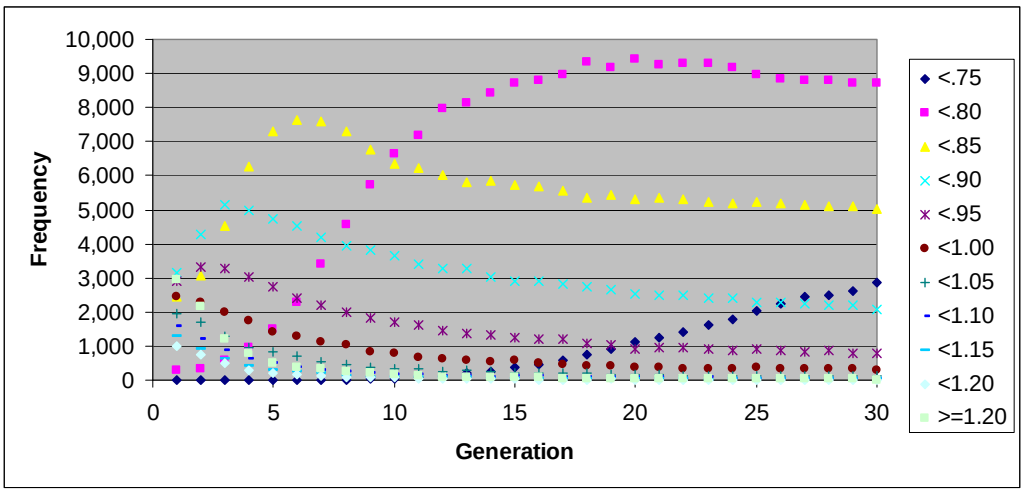


This chart shows how quickly the D-error threshold for the 5000th ranked design improves, and keeps improving through each generation. The best D-error value also improves, but

does oscillate, with the minimum value of 0.72443 in generation 27. This is substantially below the best D-error value of 0.75 in Stage 1 of the study.

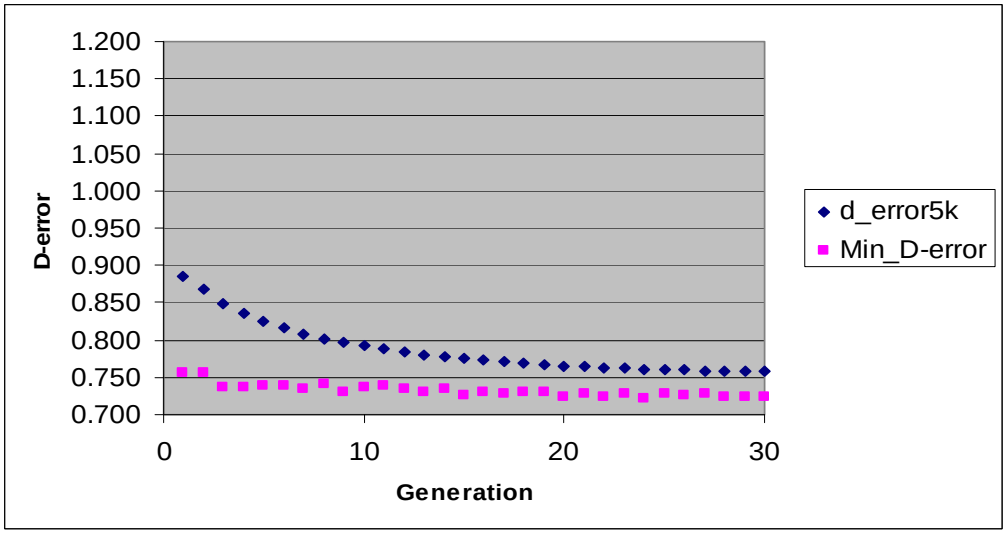
The following chart shows the progress of D-error frequency distribution starting with the best 5,000 D-error designs found in Stage 1 of the study.

Figure 9 - Experiment 1 - Initial population best designs from stage one, comparison of D-error frequency bands by generation



The two main differences from the random design starting point is that there are very few poor designs (D-error greater than 1.20) and the higher number of designs with very good D-error values, i.e., just under 3,000 with D-errors less than 0.75, which is better than the best design found in the whole of the Stage 1 study.

Figure 10 – Experiment 1 – initial population best designs from stage one, progress of best D-error and D-error of 5000th ranked design by generation



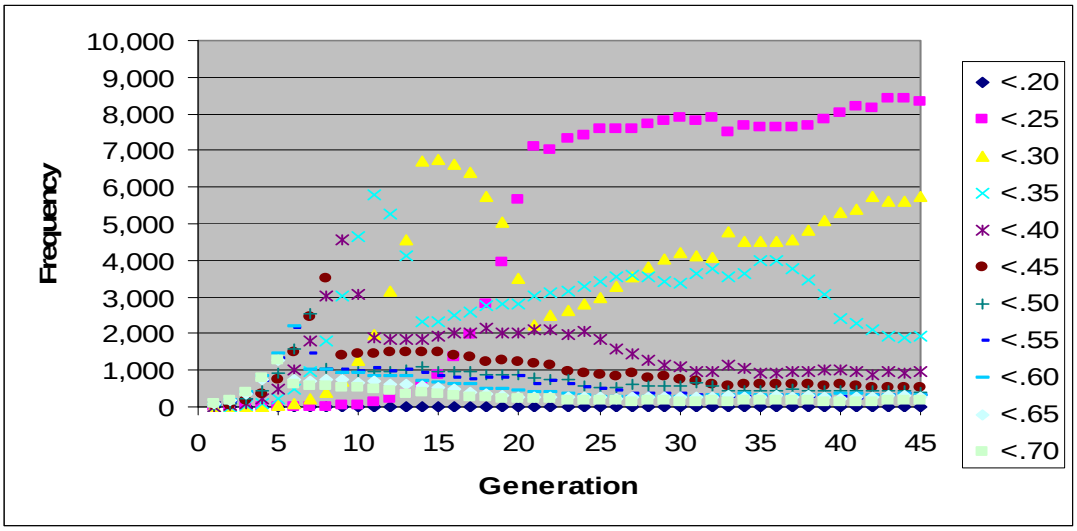
Given the better standard of designs in the initial design population, there is not the rapid decline of the D-error threshold. As with the random design starting point, the best D-error oscillates. The best D-error found was 0.72216 in generation 24, which is slightly better than that derived from 5,000 random designs. (Subsequent to composition of this paper, additional generations were run which produced a design with a D-error of 0.72107).

Of note is that it takes about 12 minutes of computing time to produce a new generation and analyse it (including the substantial overhead of writing out of all 20,000 new designs to file on a hard disk, for further exploration).

6.2 Experiment 2

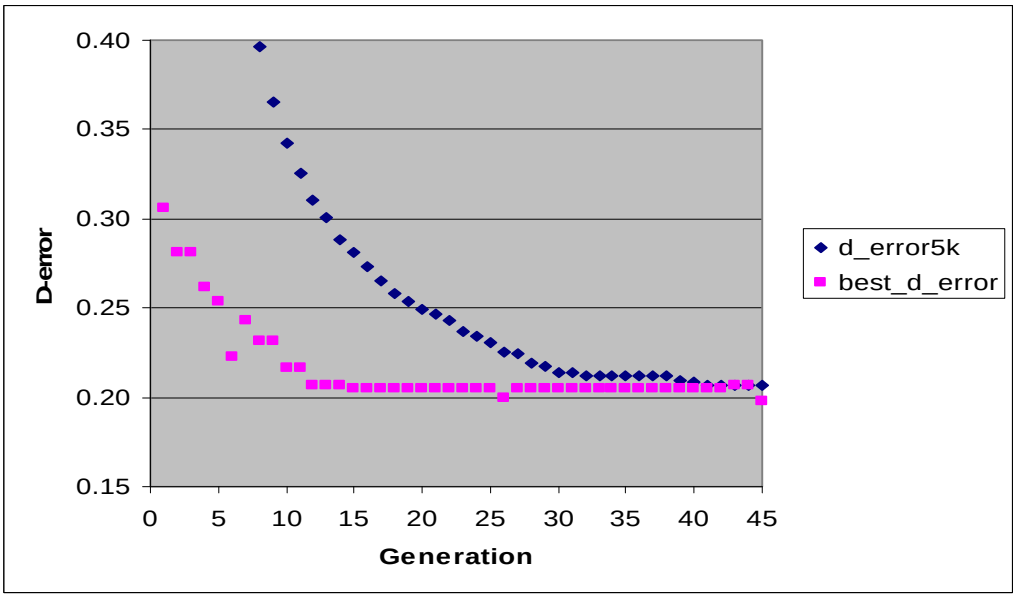
Similar charts are produced below for Experiment 2. Note that the designs with the best D-errors are in the Appendix.

Figure 11 - Experiment 2 - Initial population randomly generated, comparison of D-error frequency bands by generation



This chart shows very marked increases in the proportion of the population with good D-errors between generations 10 and 20, at which point population improvement progresses far more slowly. This is shown more clearly in Figure 12 below for the 5000th ranked design. Of note is that the best design is largely stable from generation 15.

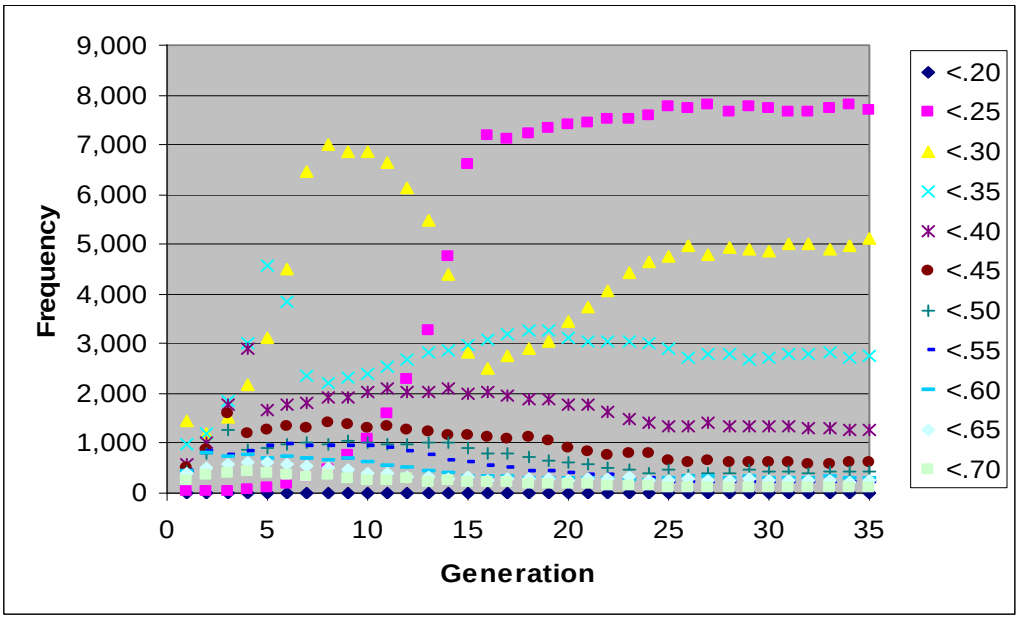
Figure 12 – Experiment 2 – initial population randomly generated, progress of best D-error and D-error of 5000th ranked design by generation



The best D-error found was 0.19803 in generation 45.

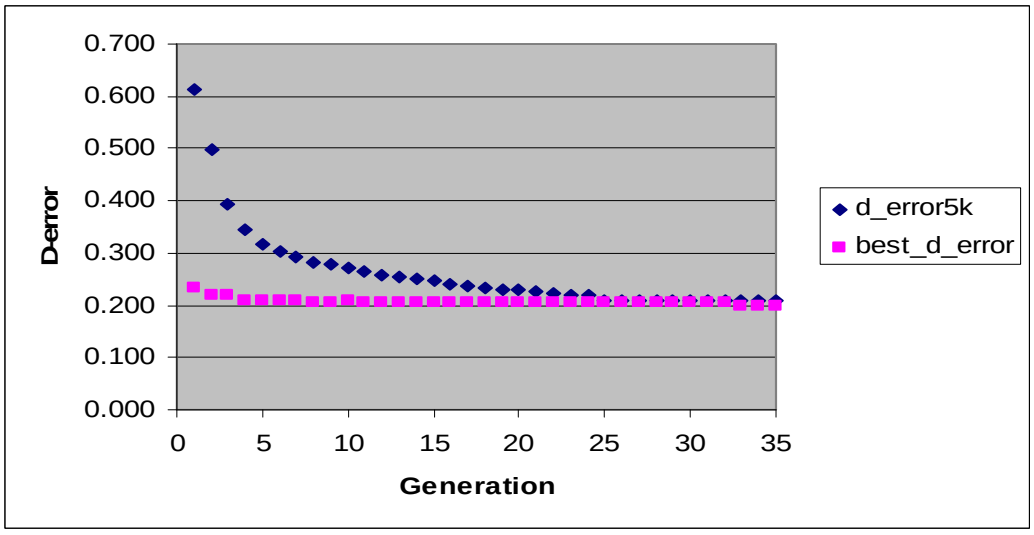
Starting with the best 5000 designs generated for Experiment 2 in stage one of the study, a similar pattern of improvement is evident in Figure 13, although it does commence earlier and stabilises around generation 15, rather than generation 20 with the randomly generated initial population.

Figure 13 - Experiment 2 - Initial population best designs from stage one, comparison of D-error frequency bands by generation



Of interest is that the best D-error reaches stability very rapidly (Figure 14), when compared with Figure 12.

Figure 14 – Experiment 2 – initial population best designs from stage one, progress of best D-error and D-error of 5000th ranked design by generation



The best D-error produced was 0.19977 in generation 33.

7 Discussion

The analysis in this paper indicates that:

- We are still no closer to being able to identify the best D-error efficient design, although we have identified a pool of very good designs for each of the two experiments.

- Designs without attribute balance appear to also have good levels of D-error efficiency.

- Genetic algorithms provide a computationally efficient method of identifying efficient experimental designs for choice experiments. This approach is certainly superior to random generation of designs, at least for the two experiments explored in this paper.

- Of the two experiments, Experiment 1, found a better design when it used the best 5,000 designs from stage one of the study, but the difference in D-error was marginal: 0.72443 (for 5,000 random designs) versus 0.72216 (for 5,000 best designs). In Experiment 2, despite the best 5,000 D-error designs being drawn from over 558 million designs, the genetic algorithm, seeded from 5,000 random designs, produced a better D-error, 0.19803 versus 0.19977.

- The computational effort required for the genetic algorithms was fractional compared with the random design strategy in Stage 1 of the study. Experiment 1 required 12 minutes to produce and analyse a generation, or a total of 6 hours for 30 generations, whilst Experiment 2 required about 30 seconds of processing time per generation, or about 23 minutes of processing time for 45 generations. This compares with about 1,500 processor hours for Experiment 1 in stage 1 and 1,200 processor hours for Experiment 2 in stage 1.

There are many dimensions along which the genetic algorithms could be altered to test for improvements in their search for better designs. These include:

- The size of the breeding population and the number of progeny per pairing;
- Considering different combination strategies (is random the optimal strategy or are there better methods?);
- Alteration strategies of attribute values – should more than a single pair of choice set/alternative attribute values be swapped?
- Selection of the best designs for breeding – should it just be the best designs? Or is there some value in throwing in some poor designs into the population to produce some 'vigour' in the population?
- The size of the pool of designs selected to breed from.

This work could explore the trade-off between getting better results and getting good results quickly.

However, such work is beyond the scope of this paper.

It is unlikely that the best D-error estimates produced in this study for Experiment 1 or Experiment 2 are the most efficient designs. Without a comprehensive frequency distribution of D-errors for each design, it is not possible to be sure that the best design has been found.

8 Conclusion

This paper demonstrates the scale of the task faced by an experiment designer when attempting to identify a good D-efficient design for use in a stated choice experiment. The

application of random generation of designs is time consuming and costly compared with the application of genetic algorithms to the task. In fact, the strength of genetic algorithms identified by the analysis reported in this paper is staggering. It is likely to be very fruitful extending this approach to mixed-logit and error components models, and incorporate Bayesian distributions of prior estimates of parameter values.

There are many dimensions along which to explore optimisation of genetic algorithms, and if the exploration in this paper is indicative, this is likely to be very fruitful.

However, the comprehensive definition of the D-error frequency distribution is likely to remain illusive for a long time yet.

9 Appendix – Best Designs produced

Table 2- Experiment 1 design with D-error 0.72216

Cset	Alt	Con 1	Con 2	A	B	C1	C2	C3
1	1	1	0	6	8	1	0	0
1	2	0	1	10	4	0	1	0
1	3	0	0	10	0	0	0	0
2	1	1	0	10	4	1	0	0
2	2	0	1	6	8	0	1	0
2	3	0	0	10	0	0	0	1
3	1	1	0	12	4	1	0	0
3	2	0	1	10	8	0	0	0
3	3	0	0	12	0	0	0	1
4	1	1	0	10	4	0	0	0
4	2	0	1	8	8	0	0	0
4	3	0	0	12	0	0	0	0
5	1	1	0	6	8	0	0	0
5	2	0	1	8	4	0	0	0
5	3	0	0	6	0	0	0	1
6	1	1	0	8	4	0	0	0
6	2	0	1	6	8	0	1	0
6	3	0	0	6	0	0	0	0
7	1	1	0	8	8	1	0	0
7	2	0	1	8	4	0	0	0
7	3	0	0	12	0	0	0	0
8	1	1	0	8	8	0	0	0
8	2	0	1	12	4	0	1	0
8	3	0	0	12	0	0	0	1

Table 3- Experiment 2 design with D-error 0.19803

Cset	Alt	Con	A	B	C
1	1	1	30	4	4
1	2	0	15	8	8
2	1	1	15	4	8
2	2	0	25	4	4
3	1	1	20	8	4
3	2	0	25	4	8
4	1	1	25	4	4
4	2	0	20	8	4
5	1	1	20	8	4
5	2	0	25	4	8
6	1	1	15	8	8
6	2	0	30	4	8
7	1	1	30	4	4
7	2	0	15	8	8
8	1	1	20	8	8
8	2	0	30	8	4

Table 4- Experiment 1 design with D-error 0.721074

Cset	Alt	Con 1	Con 2	A	B	C1	C2	C3
1	1	1	0	12	4	0	0	0
1	2	0	1	10	8	0	1	0
1	3	0	0	12	0	0	0	1
2	1	1	0	12	4	1	0	0
2	2	0	1	6	8	0	0	0
2	3	0	0	12	0	0	0	0
3	1	1	0	8	4	0	0	0
3	2	0	1	6	8	0	0	0
3	3	0	0	10	0	0	0	0
4	1	1	0	8	8	1	0	0
4	2	0	1	10	4	0	0	0
4	3	0	0	12	0	0	0	1
5	1	1	0	6	8	0	0	0
5	2	0	1	8	4	0	0	0
5	3	0	0	6	0	0	0	1
6	1	1	0	8	4	1	0	0
6	2	0	1	6	8	0	1	0
6	3	0	0	8	0	0	0	1
7	1	1	0	6	8	0	0	0
7	2	0	1	8	4	0	1	0
7	3	0	0	10	0	0	0	0
8	1	1	0	10	8	1	0	0
8	2	0	1	12	4	0	1	0
8	3	0	0	10	0	0	0	0

Acknowledgement

The author wishes to express his appreciation to the reviewers of the draft of this paper who provided very constructive and helpful comments.

References

Bliemer, MC, Rose JM and Hensher DA (2007) Constructing efficient stated choice experiments allowing for differences in error variances across subsets of alternatives *submitted to Transportation Research B,* 2007

Bliemer, MC, Rose JM and Hess S (2007) Approximation of Bayesian Efficiency in Experimental Choice Designs, *Proceedings of the 86th Transportation Research Board Annual Meeting, Washington DC,* 2007

Carlsson, F and Martinsson P (2002) Design techniques for stated preference methods in health economics, *Health Economics*, Vol 12, 2002, pp 281-294

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(3), 2007, pp 342-363

Hahn, GJ and Shapiro, SS (1966) *A Catalog and Computer Program for the Design and Analysis of Orthogonal Symmetric and Asymmetric Fractional Factorial Experiments* Schenectady, New York: General Electric Research and Development Centre

Hensher, DA Rose, JM and Greene, WH (2005) *Applied Choice Analysis – A Primer* Cambridge: Cambridge University Press

Hensher, DA Rose, JM and Bliemer MCJ (2007) Efficient Designs *Advanced Choice Experiment Design Course* Institute of Transport and Logistics Studies, Sydney: Sydney University

Huber, J and Zwerina, K (1996) The importance of utility balance in efficient choice designs *Journal of Marketing Research* XXXIII 307-317

Kanninen, BJ (2002) Optimal Design for Multinomial Choice Experiments, *Journal of Marketing Research* Vol 39 (2), 2002, pp 214-217

Louviere, JJ Hensher, DA and Swait, JD (2000) *Stated Choice Methods – Analysis and Application* Cambridge: Cambridge University Press

Rose, JM and Bliemer MCJ (2006) Designing efficient data for stated choice: Accounting for socio-demographic and contextual effects in designing stated choice experiments *11th International Conference on Travel Behaviour Research* Kyoto August 2006

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

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

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

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

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