ITLS

WORKING PAPER

ITLS-WP-09-09

Observed efficiency of a d-optimal design in an interactive agency choice experiment

By

Sean M. Puckett & John M. Rose

April 2009

ISSN 1832-570X

INSTITUTE of TRANSPORT and LOGISTICS STUDIES

The Australian Key Centre in
Transport and Logistics Management

The University of Sydney

Established under the Australian Research Council's Key Centre Program.

NUMBER: Working Paper ITLS-WP-09-09

TITLE: **Observed efficiency of a d-optimal design in an interactive agency choice experiment**

ABSTRACT:

There have been a number of recent calls within the choice literature to examine the role of social interactions upon preference formation. McFadden (2001a,b) recently stated that this area should be a high priority research agenda for choice modellers. Manski (2000) has also came to a similar conclusion and offered a plea for better data to assist in understanding the role of interactions between social agents. The interactive agency choice experiment (IACE) methodology represents a recent development in the area of discrete choice directed towards these pleas (see e.g., Brewer and Hensher 2000). The study of the influences that group interactions have upon choice bring with them not only issues that need to be overcome in terms of modelling, but also in terms of setting up the stated choice experiment itself.

Currently, the state of practice in experimental design centres on orthogonal designs (Alpizar *et al*., 2003), which are suitable when applied to surveys with a large sample size. In a stated choice experiment involving interdependent freight stakeholders in Sydney (see Hensher and Puckett 2007, Puckett *et al.* 2007, Puckett and Hensher 2008), one significant empirical constraint was difficulty in recruiting unique decision-making groups to participate. The expected relatively small sample size led us to seek an alternative experimental design. That is, we decided to construct an optimal design that utilised extant information regarding the preferences and experiences of respondents, to achieve statistically significant parameter estimates under a relatively low sample size (see Rose and Bliemer, 2006).

The D-efficient experimental design developed for the study is unique, in that it centred on the choices of interdependent respondents. Hence, the generation of the design had to account for the preferences of two distinct classes of decision makers: buyers and sellers of road freight transport. This paper discusses the process by which these (non-coincident) preferences were used to seed the generation of the experimental design, and then examines the relative power of the design through an extensive bootstrap analysis of increasingly restricted sample sizes for both decision-making classes in the sample. We demonstrate the strong potential for efficient designs to achieve empirical goals under sampling constraints, whilst identifying limitations to their power as sample size decreases.

KEY WORDS: ***Experimental design, d-optimality, interactive agency, interdependent decision making, urban freight, road user charging***

AUTHORS: Sean M. Puckett and John M. Rose

CONTACT:

Institute of Transport and Logistics Studies (C37)
The Australian Key Centre in Transport Management
The University of Sydney NSW 2006 Australia

Telephone: +61 9351 0071
Facsimile: +61 9351 0088
E-mail: itlsinfo@itls.usyd.edu.au
Internet: http://www.itls.usyd.edu.au

DATE: April 2009

1. Introduction

The paramount motivation for choosing an optimal design above an orthogonal design in a stated choice (SC) experiment is to minimise the expected standard errors in choice models that utilise the data from the experiment. This appears to be a straightforward motivation in itself; after all, why would the analyst want to induce relatively large standard errors simply as an artefact of design specification? What may be overlooked in the discussion of the merits of optimal design is the dominant force governing the choice in the first place: sample size.

There is nothing inherently wrong with orthogonal designs. Indeed, a lack of correlation across attributes in choice sets (should one be able to preserve this empirically after removing observations) is a desirable feature. Rather, orthogonal designs can require relatively large sample sizes to yield statistically significant parameter estimates in choice models. This is due to the non-linear nature of the discrete choice models, where the (co)variance matrices of such models, from which the standard errors are taken, are a function not only of the data (design), but also the choice probabilities and hence also the parameter estimates derived from the model. Whilst orthogonality relates to the correlation structure of a design, it says nothing about the choice probabilities that one is likely to obtain from models estimated using such data. Several researchers have shown that non-orthogonal designs, typically termed *efficient* designs, may produce lower standard errors than orthogonal designs for a given sample size (see e.g., Bunch et al., 1994; Bliemer and Rose, 2006; Carlsson and Martinsson, 2003; Huber and Zwerina, 1996; Kanninen, 2002; Sándor and Wedel, 2001). Thus, in cases where the large samples expected to satisfy an orthogonal design may be difficult or impossible to source due to financial, temporal or population constraints, efficient designs offer a powerful alternative.

When a choice experiment utilising an efficient design yields statistically significant parameter estimates, it is natural to assign some of the empirical success of the study to the design. Technically, however, the true power of the design is not identified simply through achieving statistical significance; after all, any given design could potentially achieve statistically significant parameter estimates despite being relatively inefficient. Fortunately, there are empirical means of identifying the robustness of experimental designs. The approach discussed in this paper, repeated bootstrapping analysis of sub-samples, is an intuitive tool for identifying the degree to which the design helped to derive the empirical results. Bootstrapping achieves this by examining the degree to which the sample size could have been further limited whilst maintaining statistical significance (and stability in behavioural implications).

In an effort to demonstrate this concept, this paper investigates the observed efficiency of a particular type of efficient design known as a *d*-efficient design that was utilised in a choice study of interdependent road freight stakeholders in Sydney, Australia. The study centred on an interactive agency choice experiment involving buyers and sellers of road freight transport services under a hypothetical variable road user charging system. An efficient design was sought due to the empirical constraints governing the experiment: difficulty in sourcing eligible respondent dyads, relatively large amounts of time needed to recruit and administer the survey for each sampled group, and relatively high expenses in administering the survey. Given the complex modelling structure that was to be applied to the choice data, these constraints made a traditional orthogonal design unlikely to produce robust behavioural results. That is, the expected sample size was not large enough to have faith in the analysts' ability to derive the desired model outputs.

Ultimately, the choice data from the *d*-efficient design utilised in the study were sufficient to obtain robust model estimates. In this paper, we examine the specific contribution of the design to the empirical results in two ways. Firstly, we analyse the model estimates that would have resulted from subsets of the sample obtained in the original study. This enables us to identify the lower limit of sample size that would have been sufficient under the experimental design given the sampled groups we were able to source. We then contrast this with a search for the smallest sample size that would have been sufficient under an orthogonal design. The juxtaposition of this information enables us to gauge the true power of the efficient design relative to an orthogonal design.

We begin the discussion with an overview of efficient designs in Section 2, and then introduce the optimal design and empirical survey utilised in the choice study in Section 3. This is followed by our empirical exercise in Section 4, and a discussion of implications for future studies in Section 5.

2. Optimal designs for multinomial logit models

The state of practice in experimental design centres on orthogonal designs (Alpizar *et al*., 2003), which are suitable when applied to surveys with a large sample size, in general. When the expected sample size for a study is small, the analyst may have reason to doubt the effectiveness of an orthogonal design. As a safeguard against yielding an insufficient sample size, the analyst may opt to develop an optimal design to achieve statistically significant parameter estimates under a relatively low sample size (see Rose and Bliemer, 2006).

An optimal design utilises extant information regarding the preferences and experiences of respondents, to specify attribute levels in choice sets that maximise the information captured when respondents select their preferred alternatives. That is, rather than setting attribute levels with respect to a constraint that they are uncorrelated across alternatives and observations as in an orthogonal design, efficient designs remove the implied assumption of equal preferences for all attributes (present in orthogonal designs) to develop alternatives that identify the preferences of respondents with greater efficiency.

Orthogonal designs ignore any extant information with respect to the preferences and experiences of respondents (i.e., marginal utility parameter estimates and attribute levels experienced in the market, respectively), yielding designs that do not achieve efficient asymptotics. That is, orthogonal designs, of which there may be any number for a given research application (with a corresponding range of efficiency that is unknown to the analyst), essentially assume that all parameters to be estimated are equal to zero, and that the attribute levels within the design are immaterial to the outcome. D-efficient designs, conversely, utilise extant information regarding the preferences and experiences of respondents, allowing for greater inferential accuracy for a given sample size, or, of paramount interest to researchers facing sampling constraints, a relatively low sample size for a given desired significance level for parameter estimates (Carlsson and Martinsson, 2003).

Other experimental design criteria can be utilised in this regard, although Kuhfeld *et al*. (1994) demonstrate that it is less difficult computationally to find a *d*-efficient design; given that the candidate efficiency criteria (e.g., *a*-efficiency and *g*-efficiency) are highly correlated with *d*-efficiency, a preference for a d-efficient design was justified in the freight study examined here (Carlsson and Martinsson, 2003). A *d*-efficient design is one of the many candidate d-efficient profiles that satisfies a desired level of statistical efficiency by minimising the expected standard errors of resulting marginal utility parameters. The *d*-efficient design utilised in the experiment was derived by specifying prior information gathered through a literature review, previous studies and a pilot study within an iterative optimisation technique[1].

[1] The optimisation technique calculated the expected d-error for randomly-generated designs under the specified prior values for attribute levels and marginal utility coefficients. The design with the lowest expected d-error was stored within the program (a macro in Microsoft Excel); whenever a new design achieved an improved error measure, the new design was stored as the preferred design. This procedure was continued until no further improvements could be found over a period deemed by the analyst to be sufficiently long to end the optimisation.

The pros and cons of *d*-efficient and orthogonal designs are compared in Table 1:

Table 1: Benefits and constraints of d-efficient and orthogonal designs

	***D*-Efficient Designs**	*Orthogonal Designs*
Required Sample Size to Achieve Desired Standard Errors	Generally small	Generally large
Statistical Knowledge Required	Relatively large	Relatively small
Prevalence in the Literature	Not utilised often	Predominant design form
Ease of Design Generation	Designed through software or first principles	Designed through software, first principles, websites and published arrays
Evidence of Priors Required	Yes	No
Flexibility in Generation	User defines the constraints of the design	Orthogonality is constrained by the number of alternatives, attributes and their levels

Table 1 highlights the primary reasons that orthogonal designs are the most common choice in experimental design applications: orthogonal designs have relatively low user-input requirements, can be generated by using many readily available sources, and their use is widely accepted in the literature. Indeed, if the expected sample size is reasonably large, it is fair to expect an orthogonal design to lead to sufficiently small standard errors for model outputs. However, if the expected sample size is not large, the additional knowledge and resources required to derive a *d*-efficient design need not be prohibitive. The flexibility gained in removing the constraints relating to orthogonality is powerful, and may offset (perhaps greatly) the burden associated with establishing prior information and generating the design.

When working on the experimental design in our application, we expected the empirical framework to be relatively complex. Indeed, the empirical models were generalised mixed logit models. However, it should be noted that the experimental design was calibrated using a relatively simple multinomial logit (MNL) framework. At the time of the empirical study, the asymptotic variance-covariance matrix for the mixed logit model was unknown. The only available means for deriving a d-efficient design based upon a mixed logit model would have been simulation, which would have taken prohibitively long to carry out; a simulation of each iteration of the mixed logit model for each design to test would have been required, which was not a practical option. Fortunately, Bliemer and Rose (2008, 2009) have demonstrated that the MNL model offers a reasonable approximation of the mixed logit model when generating optimal designs.

Following the exposition of Carlsson and Martinsson (2003) within an MNL framework, when developing a *d*-efficient design for a choice experiment involving a revealed-preference-based reference alternative and SC alternatives, the analyst's task is to estimate, with the highest degree of precision feasible, the parameters of the utility functions for the reference alternative *r* and the SC alternatives *s*, respectively:

$$U_{ir} = \alpha + \boldsymbol{\beta_k} \boldsymbol{x_k} + \varepsilon \qquad (1)$$

$$U_{is} = \boldsymbol{\beta_k} \boldsymbol{x_k} + \varepsilon \qquad (2)$$

where α represents an alternative-specific constant representing the real-market nature of the reference alternative, $\boldsymbol{\beta_k}$ represents the vector of desired marginal utility parameter estimates, $\boldsymbol{x_k}$ represents the levels of the corresponding vector of *k* attributes in the alternative, and ε represents the unobserved effects, which are assumed to be independently and identically distributed type I extreme value. The fundamental difference between the two utility functions in the freight study is that the attribute level for variable road-user charges (VUCs) in the reference alternative is always equal to zero.

As demonstrated by McFadden (1974), the covariance matrix of the expected maximum likelihood estimators (i.e., those based upon prior information) is a function of the observed marginal utilities of respondents,

$$\Omega = \left[\sum_{n=1}^{N} \sum_{j=1}^{J} z^{'}_{jn} * P_{jn} * z^{-1}{}_{jn} \right] \tag{3}$$

where

$$z_{jn} = x_{jn} - \sum_{i=1}^{J_n} x_{in} * P_{in} \tag{4}$$

and where x_{jn} represents the vector of attribute levels for an alternative j (numbered 1 to J) in choice set n (numbered 1 to N), and P_{jn} represents the choice probability for alternative j in choice set n (Carlsson and Martinsson, 2003).

This is intuitive, as the choice probabilities that are observed are a direct function of the preferences underlying the choices made. Therefore, the covariance matrix Ω is a function of both the marginal utilities of respondents, which are invariant across alternatives, and the attribute level combinations corresponding to a set of alternatives on offer, which, in the case of the SC alternatives, are under the control of the analyst.

A *d*-efficient design is found when maximising the inverse of the determinant of Ω (scaled by an exponent incorporating the number of parameters to estimate *K*), which Kanninen (2002) points out is the (scaled) expected value of the Hessian of the log likelihood function, multiplied by -1:

$$\textbf{max} \left[\left| \Omega \right|^{1/K} \right]^{-1} \tag{5}$$

Importantly, Kanninen clarifies that, due to the central role of the covariance matrix within the search for *d*-efficiency, by maximising *d*-efficiency (should the priors be correct), one minimises the magnitude of the asymptotic confidence region around the parameter estimates. Hence, the efficiency of the design is critically dependent upon the manner in which the attribute levels are specified for each alternative. Consequently, by utilising prior information about the likely preferences of respondents and the likely reference attribute levels they would specify, greater efficiency can be achieved through minimisation of the estimated covariances by manipulating the combination of attribute levels on offer across alternatives for each choice set (Huber and Zwerina, 1996; Carlsson and Martinsson, 2003).

3. Freight survey and its *d*-efficient design

3.1 Freight survey

The primary focus of the freight study was to establish how the implementation of a variable user charging (VUC) system may affect both the levels of service offered by the traffic infrastructure and the costs of transporting freight. The SC experiment was designed to capture the preferences of freight stakeholders under the range of levels of services and costs that may be present under a distance-based road user charging system, each of whom interacts as described above. Respondents from freight firms and their clients were asked to choose from among a set of alternative urban freight trip options for a particular consignment, each of which contained its unique mix of levels of service components and cost. These alternatives represent potential means of coping with a (hypothetical) VUC system. Each of these alternatives was framed in reference to recent experience that forms the basis for the SC alternatives; i.e., a recent goods movement, its corresponding performance measures, and indicators of the relationship between the two firms involved.

The primary desired estimation outputs were the sensitivities of buyers and sellers of urban freight services to trade-offs between elements of travel time and cost under VUCs. These include two measures of the value of time savings: one for travel time in free-flow traffic conditions (i.e., those where the truck can travel at the speed limit and manoeuvre without difficulty), and one for travel time in slowed-down conditions (i.e., those where the truck has difficulty in travelling at the speed limit, and where manoeuvring is impeded by the level of other vehicles present). The other two temporal measures are related to transactions time and reliability in arrival time. The former is measured as the value of waiting time savings (i.e., time spent waiting at delivery destinations whilst unable to unload goods due to queuing); the latter is measured as the value of reliability gains (i.e., percentage increases in the probability that a truck will reach its delivery destination without incurring a penalty due to missing a specified arrival time window).

Respondents were asked to assume that, for each of the choice sets given, the same cargo needs to be carried out for the same client discussed earlier in the survey, subject to the same constraints faced when the reference trip was undertaken. Respondents were then informed that the choice sets involve three alternative methods of carrying out the trip: their experienced trip and two SC alternatives that involve VUCs. The choice tasks were described to respondents as involving two steps. The first step is to indicate which alternatives would be preferable if the two organisations had to reach agreement, whilst the second step is to indicate what information mattered when making each choice.

Respondents were faced with four choice sets if representing a freight firm and eight choice sets if representing a client of a freight firm. The difference is attributable to the relatively larger burden placed on respondents from freight firms, in that they must supply the trip- and relationship-specific details required to establish the choice setting and reference alternative. The exact four choice sets answered by a given respondent from a freight firm are given to the corresponding sampled client. The additional four choice sets faced by the sampled client use the same reference alternative as the other four choice sets.

The attributes within each choice set are: free-flow travel time, slowed-down travel time, total time waiting to unload goods, likelihood of on-time arrival, freight rate paid by the client, fuel cost, and variable charges, the latter of which always takes a value of zero for the reference alternative. Each of these attributes except for the freight rate (which is not a design attribute) are either an input into a road-user charging policy (i.e., changes in fuel taxes, road user charges), or direct functions of such a policy. The likelihood of on-time arrival was chosen in preference to other measures of reliability or travel time variability, as in-depth interviews revealed that on-time arrival rates (defined within the experiment as the likelihood of reaching the delivery destination(s) close enough to the time window agreed upon to avoid being penalised) are a key measure of reliability.

The levels and ranges of the attributes were chosen to reflect a range of available routing and scheduling options under a hypothetical VUC system. The reference alternative was utilised to offer a base, around which the stated choice design levels were pivoted. The resulting mixes represent alternatives including: taking the same route at the same time as in the reference alternative under new traffic conditions, costs, or both; and taking alternative, previously less-favourable routes, departing at alternative, previously less-favourable times, or both, with corresponding levels of traffic conditions and costs.

In all cases except for the VUC's, referred to in the SC experiment as a distance-based charge, the attribute levels for each of the SC alternatives are expressed as deviations from the reference level, which is the exact value specified in the corresponding non-SC questions, unless noted:

(1) Free-flow time: –50%, –25%, 0, +25%, +50%
(2) Slowed-down time: –50%, –25%, 0, +25%, +50%
(3) Waiting time at destination: –50%, –25%, 0, +25%, +50%
(4) Probability of on-time arrival: –50%, –25%, 0, +25%, +50%, with the resulting value rounded to the nearest 5% (e.g., a reference value of 75% reduced by 50% would yield a raw figure of 37.5%, which would be rounded to 40%). If the resulting value is 100%, the value is expressed as 99%. If the reference level is greater than 92%, the pivot base is set to 92%. If the pivot base is greater than 66% (i.e., if 1 ½ times the base would be greater than 100%) let the pivot base equal X, and let the difference between 99% and X equal Y. The range of attribute levels for on-time arrival when X > 66% are (in percentage terms): X–Y, X–.5*Y, X, X+.5*Y, X+Y. This yields five equally-spaced attribute levels between X–Y and 99%.
(5) Fuel cost: –50%, –25%, 0, +25%, +50% (representing changes in fuel taxes of –100%, –50%, 0, +50%, +100%). Note: Fuel taxes represented approximately half of fuel prices in Australia at the time of the study.
(6) Distance-based (or variable user) charges: Pivot base equals one-half of the reference fuel cost, to reflect the amount of fuel taxes paid in the reference alternative. Variations around the pivot base are: –50%, –25%, 0, +25%, +50%

One potential complication that we identified is that changes in levels of service and operating costs could lead to upward or downward adjustments in the freight rate charged by the transport company. Although obvious, incorporating an endogenous (at least to the freight transport provider) choice that could swamp the changes in costs into the experimental design is not a simple matter. To accommodate this, we developed a method to internalise this endogeneity and uncertainty, making it exogenous to the final choice. For each SC alternative involving a net change in direct operating costs (i.e., a decrease in fuel costs that is not equal in magnitude to the value of the new distance-based charges), respondents from freight firms are asked to indicate by how much of the net change in costs they would like to adjust their freight rate (Fig. 5). Hence, the freight rate, which is not a design alternative, yet is clearly an important contextual effect, is allowed to vary across SC alternatives under changes in net operating costs. The specific range over which the freight rate may vary is bounded by the change in net operating cost for each alternative.

3.2 D-efficient design

As mentioned in Section 3.1, each choice set in the study contained a respondent-specified reference alternative, along with two SC design alternatives. Complicating the design generation process was the fact that the full design is partitioned into two groups of choice sets. Respondents from freight firms, who were hypothesised to have one set of marginal utilities, were given four choice sets; respondents from shippers, who were hypothesised to have a set of marginal utilities that differed to those of freight firm respondents, were given the identical four choice sets that were given to a corresponding freight firm respondent, along with four unique choice sets.

When generating the experimental design, it was necessary to specify appropriate prior values for marginal utility parameters and attribute levels. The survey pre-testing phase and literature reviews identified a range of plausible prior specifications across the two respondent classes, to accommodate for the likely divergent preferences for transporters and shippers. The prior parameter estimates for the design are shown in Table 2:

Table 2: Prior parameter values for d-efficient design

	Transporters	**Shippers**
Free-flow Time	**-0.047**	**-0.024**
Slowed-down Time	**-0.066**	**-0.024**
Waiting Time	**-0.057**	**-0.024**
Probability of On-Time Arrival	**0.038**	**0.038**
Fuel Cost	**-0.058**	**-0.029**
Distance-Based Charges	**-0.116**	**-0.058**

Parameter estimates for free-flow travel time and fuel cost for freight transport operators from a previous study at the Institute of Transport and Logistics Studies were used as priors for the same attributes in the design for respondents from freight firms. The prior for likelihood of on-time arrival was specified as the negative value of the prior for fuel cost, yielding a prior value of reliability gains of A$1 per percentage point increase in reliability, using the cost of fuel as a base measure. This figure was selected as a hedge between relatively lower and higher priors available. The parameter value for variability in travel time from the aforementioned study was a candidate prior (after being multiplied by negative one due to the inverse behavioural relationship between the two concepts of reliability and variability). The resulting prior value of reliability gains using this measure was approximately A$0.65 per percentage point increase in reliability.

Priors for the remaining parameters for respondents from freight transport providers were developed using the following heuristics. Firstly, the value for slowed-down time was found by multiplying the prior for free-flow time by 1.4, which is a ratio supported by previous travel studies at ITLS. The prior for waiting time was set as a weighted average of the free-flow and slowed-down priors (two-thirds of the former and one-third of the latter). Lastly, the prior for distance-based charges was set as two times the prior for fuel cost, to account for scaling effects in attribute values; that is, the average attribute values for fuel cost are expected to be twice as high as those for distance-based charges, and hence a base assumption of equivalent aversion to both cost measures necessitates scaling the parameter for distance-based charges.

These priors were adjusted for respondents from clients of freight transport providers using the following heuristics. Firstly, the value for free-flow time was specified as one-half the value of the prior for free-flow time for freight transporters. A value of one-half of the corresponding prior for freight transporters was chosen as a parsimonious hedge amongst uncertainty in which the plausible prior value ranged between zero and the value held by freight transporters. Secondly, we assumed no variation in preferences for types of travel time, and hence set the priors for slowed-down time and waiting time equal to this value. The priors for cost measures were set equal to the corresponding priors for freight transporters, due to the ability of freight transporters to pass along the new costs to shippers within the experiment. Lastly, the prior for reliability was set as equal to the corresponding prior for freight transporters, because shippers value reliability, and hence the established prior was the best value available for us to utilise.

Once the prior parameter values were established, we needed to identify appropriate prior values for the corresponding attribute levels. Whilst it was known the attribute levels would be the same for transporters and shippers, we needed to establish whether one prior attribute level would be sufficient for each attribute in the design, or whether it was preferable to segment the design into multiple classes of trips.

We chose to generate separate the design into two segments: those involving trips of less than two hours, and those lasting two to seven hours. The motivation for this segmentation arose from the data source utilised to set these priors. Global positioning systems (GPS) devices were placed in four freight vehicles operating for a major freight transport company in Sydney for one week. The data from the GPS devices was used to measure distances and times for freight delivery tours for the vehicles. Approximately half of the trips measured took two hours or less to complete, whilst virtually all of the remainder fell within the seven-hour limit established for the choice experiment.

Table 3 shows the prior attribute levels established for two broad trip-length segments, trips of two hours or less, and trips of two to seven hours:

Table 3: Prior attribute levels for the d-efficient design

	Trips Less than Two Hours	**Trips Greater than Two Hours**
Free-flow Time	**40 minutes**	**140 minutes**
Slowed-down Time	**20 minutes**	**45 minutes**
Waiting Time	**20 minutes**	**45 minutes**
Probability of On-Time Arrival	**75 percent**	**75 percent**
Fuel Cost	**$11.00**	**$30.00**
Distance-based Charges	**$5.50**	**$15.00**

The average travel time for each trip length segment was divided into two, with one half specified as the prior free-flow time and the other half specified as the prior slowed-down time. With no further information on the likely proportion of slowed-down time in total travel time, an even split was determined to be the most parsimonious decision. The minimum value of time spent unloading at a destination was used as the baseline for unloading time; this value was deducted from the average time spent at delivery destinations, with the difference multiplied by the average number of deliveries made in each segment to find the prior value for waiting time. The prior value for likelihood of on-time arrival was established as the arithmetic average of on-time arrival rates for primary and secondary retail freight, as revealed by the in-depth interviews. The prior value for fuel cost was established by multiplying the average fuel efficiency of the predominant vehicle type (rigid truck) by the average distance travelled within each trip length segment, and multiplied by the current price of diesel fuel. This yielded a base value for distance-based charges in the design, equal to one half of the prior for fuel cost.

The pilot data did not reveal significant parameter estimates when analysed within a basic multinomial logit model, and hence the prior parameter estimates were not amended as a result of the pilot. However, the pilot confirmed the presence of meaningful trade-offs within choice sets, in that the reference alternative did not dominate the SC alternatives. Likewise, the pilot confirmed the merit of the prior attribute levels utilised within the experimental design. That is, the observed RP data offered by respondents were consistent with the prior attribute levels.

The *d*-efficient design ultimately utilised in this research was found through the use of a search algorithm designed at ITLS, which was adjusted for the complex nature of the interactive agency survey. The search algorithm was designed to accommodate: (1) the presence of a mix of a reference alternative and two SC alternatives that are generic to one another; (2) the constraint that the four choice sets faced by a respondent from a freight firm must also be given to a respondent to a corresponding sampled client; and (3) hypothesised preference heterogeneity across agent types.

Whilst extensive iteration of the algorithm could not guarantee that *d*-efficiency was maximised globally, the design selected by the algorithm achieved superior *d*-efficiency to the other designs generated in the search process. The final *d*-efficient design utilised in this research is given in the Appendix. This design was selected as the design most capable of achieving precise parameter estimates, conditional on the prior values specified.

4. Empirical analysis

The freight study yielded a sample of 145 transporters (for a total of 580 choice observations) and 138 shippers (for a total of 1106 choice observations). The analysis in this section centres on repeated bootstrapping procedures that were conducted within the software program Ngene. Utilising this software, we drew repeated sub-samples of transporters and shippers at increasingly restricted sample sizes. For each sub-sample, we re-estimated base empirical models of transporter and shipper preferences with respect to the choice sets faced within the freight study.

The use of repeated sub-samples for each target sample size allows us to identify trends in the mean and standard deviation of model outputs as the sample size decreases. With this information we can gauge whether the experimental design used in the study would have been sufficient to achieve the desired statistical significance under a given restricted sample size. Likewise, trends in the mean and standard deviation of the model outputs across restricted sample sizes enable us to identify whether there is a potential for model outputs to be biased or unreliable at a given sample size.

4.1 Results for transporters

We begin the analysis by estimating the most complete replica of the base empirical model for transporters in the original freight study, with three key differences. Firstly, we estimated a model that includes every transporter choice observation, rather than excluding potential outliers as in the original study; this gives us the largest possible sample size to use as a benchmark when comparing sub-sample sizes. Secondly, we only tested unconstrained normal and uniform distributions for possible effects of unobserved preference heterogeneity; this restriction is due to a limitation in the software utilised in the bootstrapping exercise. Thirdly, we isolated transporters for their own model rather than pooling transporters and shippers into one model; this was done to avoid sub-samples with implausible membership (e.g., shippers whose corresponding transporter partner are not included in the sample).

Given these constraints, we settled upon a base model for transporters with each of the attribute constructs in the original transporter model (travel time, probability of on-time arrival, fuel cost, freight rate, distance-based charges), plus one interaction term that was in the original model (free-flow travel time multiplied by distance travelled), which serves to isolate separate disutilities for free-flow travel time and slowed-down travel time. The probability of on-time arrival, freight rate and distance-based charges are modelled as normally-distributed random parameters.

Table 4A highlights both the statistical significance of the explanatory variables and the degree of misspecification for each of the prior parameter values (expressed as marginal rates of substitution with respect to fuel cost):

Table 4A: Transporter base model versus priors

(t-statistics in parentheses; all random parameters distributed normally)

	Full Sample	**Marginal Rates of Substitution* (MRS)**	**Prior MRS***
Parameter			
Reference Alternative	0.861 (3.209)		
Free-Flow Time (min)	-0.009 (-2.284, 1.669**)	0.819	0.810
Slowed-Down Time (min)	-0.015 (-2.284)	1.364	1.138
Probability of On-Time Arrival (%) - Mean	0.030 (1.985)	-2.727	-1
Probability of On-Time Arrival (%) – Std. Dev.	0.056 (2.07)		
Fuel Cost ($)	-0.011 (-2.383)	1	1
Freight Rate ($) - Mean	0.002 (0.354)	-0.182	--
Freight Rate ($) – Std. Dev.	0.008 (2.213)		
Distance-Based Charges ($) - Mean	-0.008 (-1.377)	0.727	2
Distance-Based Charges ($) – Std. Dev.	0.011 (2.344)		
Waiting Time (min)	0 (Not in model)	0	0.983

****-Marginal rates of substitution are for mean values with respect to the mean value for fuel cost.***

*****-t-statistic for an interaction term involving distance travelled; the mean effect of this interaction is included in the parameter for free-flow time for comparison with the prior***

Each of the linear effects without a random specification shows a high level of statistical significance, with the appropriate sign. Transporters show a strong preference for the reference alternative (which has no distance-based charging component), and disutility for travel time and fuel cost. The mean and standard deviation for on-time arrival probability enters the model highly significantly, demonstrating both a general preference for reliable travel and heterogeneity in the degree to which this is preferred. The freight rate shows an insignificant positive marginal utility, which confirms a hypothesis of rational passing of distance-based charges and fuel costs along to shippers (i.e., once transporters decide how much of the changes in their costs to pass along to customers via the freight rate, they are indifferent to the amount they are compensated across alternatives). However, there is a highly significant degree of preference heterogeneity with respect to the freight rate, suggesting that there are some transporters who are much more sensitive to the freight rate than others; hence, the low mean utility could also be related to other effects. Lastly, distance-based charges have a mean negative effect on utility less than the disutility of fuel cost, although this effect has only minor statistical significance. Transporters appear to have strongly varying sensitivities to distance-based charges, with a highly significant standard deviation of marginal disutilities for the charges.

Some relationships between these parameter estimates are close to those amongst the assumed prior parameter values, whilst others reveal relative sensitivities that differ strongly from the priors. Most notably, the mean estimated marginal rates of substitution for both free-flow and slowed-down travel time with respect to fuel cost are quite similar to their corresponding prior values, at 0.82 dollars per minute in free-flow conditions versus a prior value of 0.81 and 1.36 dollars per minute in slowed-down conditions versus a prior value of 1.14. Conversely, sampled transporters demonstrated mean sensitivities to on-time arrival reliability and distance-based charges with respect to fuel cost that were highly divergent from the assumed prior relationships, at –2.73 dollars per percentage point of reliability versus a prior value of 1 and 0.73 dollar spent on fuel per dollar spent on distance-based charges versus a prior value of 2.

We can also compare the assumed prior attribute levels for the reference alternative versus those observed in the study (which, unlike the preference estimates are constant across decision-maker class). In contrast to the relatively large discrepancies with respect to prior and observed marginal utility parameters, Table 4B shows that the observed mean attribute levels were generally similar to the prior values:

Table 4B: Observed mean attribute levels by segment versus priors

	Trips Less than Two Hours		Trips Greater than Two Hours	
	Assumed Priors	Observed Mean	Assumed Priors	Observed Mean
Free-flow Time (min)	**40**	**50.39**	**140**	**234.68**
Slowed-down Time (min)	**20**	**26.49**	**45**	**51.02**
Waiting Time (min)	**20**	**34.07**	**45**	**64.58**
Probability of On-Time Arrival (%)	**75**	**82.54**	**75**	**81.54**
Fuel Cost	**$11.00**	**$31.69**	**$30.00**	**$237.99**

In the two-hour-or-less segment, each of the observed travel time and on-time arrival values are similar to the priors, with free-flow travel time, slowed-down travel time and waiting time only around 10, 6 and 14 minutes greater than the prior values, respectively. Likewise, the mean on-time arrival probability is only around eight percent (in magnitude) greater than the prior value. The observed cost measures are much higher, however, with levels almost three times as large as the priors.

In the over-two-hours segment, the priors reflect a mean trip involving considerably less free-flow time and distance travelled than observed in the sample. The observed free-flow time and costs are much higher than the prior values. Still, the observed values for slowed-down time, waiting time and on-time arrival probability are relatively close to the priors, with discrepancies of only around six minutes, 20 minutes, and six percent, respectively.

Considering the marginal utility parameters and the mean reference alternative attribute levels together, the picture that emerges is that the prior specification was subject to some considerable error and uncertainty that was more pronounced for some attributes. Despite the degree of prior misspecification, which could be expected in many choice studies (after all, why should we collect the data if we already knew the answers we were searching for?), the presence of statistically significant parameter estimates under a small sample and smaller sub-samples (to follow) confirms the power of optimal designs.

We now turn to the bootstrapping exercise for transporters. We re-estimated the base model for 100 randomly-selected sub-samples for restricted sample sizes (i.e., cohort sizes) of decreasing multiples of ten. Beginning with a cohort size of 140 (the closest multiple of ten below the full sample size of 145) we estimated repeated sub-samples on increasingly smaller cohort sizes to gauge the strength of the experimental design. Table 5 shows the relative stability in mean parameter estimates as the sample size is increasingly restricted:

Table 5: Mean bootstrap parameter estimates by cohort size

(Transporters, 100 Sub-samples per Cohort Size, 4 Observations per Respondent)

	All (145)	**140**	**130**	**120**	**110**	**100**	**90**	**80**	**70**
Reference Alternative	0.8614	0.8662	0.8560	0.8563	0.8458	0.8586	0.8473	0.8661	0.8385
Travel Time (min)	-0.0146	-0.0145	-0.0145	-0.0144	-0.0147	-0.0146	-0.0149	-0.0147	-0.0148
Probability of On-Time Arrival (%) - Mean	0.0304	0.0305	0.0309	0.0312	0.0310	0.0310	0.0308	0.0307	0.0299
Probability of On-Time Arrival (%) – Std. Dev.	0.0560	0.0560	0.0557	0.0561	0.0547	0.0550	0.0542	0.0545	0.0534
Fuel Cost ($)	-0.0105	-0.0106	-0.0106	-0.0105	-0.0105	-0.0105	-0.0105	-0.0106	-0.0107
Freight Rate ($) - Mean	0.0015	0.0017	0.0017	0.0015	0.0014	0.0013	0.0013	0.0014	0.0014
Freight Rate ($) – Std. Dev.	0.0077	0.0077	0.0077	0.0078	0.0078	0.0077	0.0076	0.0075	0.0074
Free-Flow Time (min) * Distance (in '000 km)	0.0239	0.0238	0.0241	0.0237	0.0242	0.0241	0.0249	0.0240	0.0241
Distance-Based Charges ($) - Mean	-0.0080	-0.0080	-0.0082	-0.0082	-0.0083	-0.0081	-0.0085	-0.0083	-0.0094
Distance-Based Charges ($) – Std. Dev.	0.0111	0.0111	0.0112	0.0115	0.0115	0.0114	0.0117	0.0111	0.0121

The mean parameter estimates appear fairly stable as the cohort size falls, implying no general tendency toward bias under small samples. That is, other than some small fluctuations in mean values as the sample size is reduced, a relatively small sample size would have given similar results as the full sample, on average. Indeed, the mean parameter values show no general misbehaviour until the cohort size falls below 80, which approaches only one-half of the sample.

The mean parameter estimates across cohort sizes do not tell the whole story, however. In practice, under any given restricted sample size the analyst would only have one sample with which to work. One question that follows the first test for consistency directly is whether the parameter estimates that one would obtain under a restricted sample size would tend to be statistically significant. Table 6 compares the frequencies (out of 100) with which each parameter estimate meets the standard of a p-value below .05 and .1, respectively, as the sample size is reduced:

Table 6: Frequency of parameter estimates significant at target confidence levels

by cohort size (transporters)

Frequency of Significance at the 95% Confidence Level								
				Cohort Size				
Attribute	**140**	**130**	**120**	**110**	**100**	**90**	**80**	**70**
Reference Alternative	100	100	100	99	91	86	80	61
Travel Time (min)	100	87	69	59	38	30	14	11
Probability of On-Time Arrival (%) - Mean	52	41	25	21	15	12	6	2
Probability of On-Time Arrival (%) – Std. Dev.	72	51	35	16	13	12	5	3
Fuel Cost ($)	100	99	90	72	46	28	15	7
Freight Rate ($) - Mean	0	0	0	0	0	0	0	0
Freight Rate ($) – Std. Dev.	96	77	60	45	37	25	13	8
Free-Flow Time (min) * Distance (in '000 km)	0	0	0	1	5	3	4	1
Distance-Based Charges ($) - Mean	0	0	0	0	0	1	0	0
Distance-Based Charges ($) – Std. Dev.	99	90	81	60	47	32	10	11

Frequency of Significance at the 90% Confidence Level								
	Cohort Size							
Attribute	**140**	**130**	**120**	**110**	**100**	**90**	**80**	**70**
Reference Alternative	100	100	100	99	96	95	93	84
Travel Time (min)	100	100	100	97	81	74	54	39
Probability of On-Time Arrival (%) - Mean	100	91	82	67	51	38	26	14
Probability of On-Time Arrival (%) – Std. Dev.	100	93	86	74	55	42	32	18
Fuel Cost ($)	100	100	100	95	89	79	57	40
Freight Rate ($) - Mean	0	0	0	0	0	0	0	0
Freight Rate ($) – Std. Dev.	98	93	84	72	60	51	42	27
Free-Flow Time (min) * Distance (in ‘000 km)	40	37	19	21	19	22	10	11
Distance-Based Charges ($) - Mean	0	1	0	3	1	4	1	2
Distance-Based Charges ($) – Std. Dev.	100	99	97	91	81	69	56	40

A preference for the reference alternative was the strongest effect statistically in the base model, and hence serves as a useful initial gauge of the relative strength of the experimental design at each cohort size. The reference alternative has a statistically significant at the 95 percent confidence level in every sub-sample, save one outlier, through a sample size of 110; the reference alternative maintained a statistically significant effect at the 90% level consistently in sample sizes as low as 80. Given that the reference alternative was the strongest explanatory variable in the base model, it would follow that other explanatory variables would require larger sample sizes to reach a desired significance level. Indeed, travel time has difficulty in achieving a p-value below .05 under mildly-restricted sample sizes; travel time does maintain a p-value below .1 in almost all cases down to a sample size of 110. On-time arrival probability appears to require more choice observations than travel time to achieve a given p-value, although its status as a random parameter may contribute to this. Fuel cost performs well under relatively sample sizes, generally maintaining statistical significance at the 95% and 90% levels at sample sizes of 120 and 100, respectively. The interaction between free-flow travel time and trip distance loses significance quickly as the sample size is reduced, implying that at lower sample sizes a different modelling construct would be needed to identify separate effects for free-flow and slowed-down travel time. Lastly, the mean disutility of distance-based charges is insignificant at all restricted sample sizes, but heterogeneity around the mean is generally statistically significant at the 95% and 90% confidence levels at sample sizes of 130 and 110, respectively.

Now that we have looked into the mean values of each marginal utility estimate and the relative tendencies for each parameter estimate to be statistically significant at increasingly restricted sample sizes, it is important to consider the degree to which estimates from any given sub-sample would tend to differ from the full sample value. It is possible to observe repeated sub-samples that yield mean parameter estimates that are both close to the full sample values and which have a high probability of being statistically significant, yet display the potential to include individual sub-samples with parameter estimates that are different from the mean value *and* are statistically significant, which is problematic. Table 7 shows the normalised standard deviation of each parameter estimate by cohort size. This gives us a scaled indicator of how closely grouped each parameter estimate is across repeated sub-samples:

Table 7: Standard deviation of parameter estimates as a percentage of the mean by cohort size (transporters)

	140	**130**	**120**	**110**	**100**	**90**	**80**	**70**
Reference Alternative	0.035	0.074	0.112	0.150	0.185	0.179	0.200	0.246
Travel Time (min)	-0.046	-0.079	-0.093	-0.121	-0.154	-0.177	-0.195	-0.215
Probability of On-Time Arrival (%) - Mean	0.057	0.110	0.148	0.187	0.209	0.232	0.265	0.317
Probability of On-Time Arrival (%) – Std. Dev.	0.057	0.096	0.134	0.175	0.197	0.246	0.277	0.307
Fuel Cost ($)	-0.052	-0.095	-0.124	-0.167	-0.187	-0.252	-0.271	-0.327
Freight Rate ($) - Mean	0.424	0.779	1.151	1.522	1.870	2.473	2.560	3.193
Freight Rate ($) – Std. Dev.	0.098	0.183	0.239	0.305	0.367	0.403	0.484	0.544
Free-Flow Time (min) * Distance (in ‘000 km)	0.070	0.121	0.141	0.188	0.243	0.308	0.336	0.365
Distance-Based Charges ($) - Mean	-0.100	-0.193	-0.221	-0.325	-0.406	-0.466	-0.526	-0.600
Distance-Based Charges ($) – Std. Dev.	0.054	0.108	0.154	0.202	0.298	0.289	0.359	0.428

Under an assumption of normality, we would expect a little more than 95 percent of the parameter estimates to lie within a range of two times the normalised standard deviation away from the mean. Beginning with the marginal utility of the reference alternative, a little more than 95 percent of the parameter estimates lie within seven percent of the mean value at a restricted sample size of 140. This range grows quickly as the sample size is reduced, doubling then the sample size is restricted to 130 (indeed, each of the parameter estimate distribution ranges from the mean roughly doubles when the sample size is reduced from 140 to 130). This degree of uncertainty is stepped up incrementally again when restricting the sample size to 120 (the lowest sample size at which one may be reasonably confident in achieving strong statistical significance throughout the model), at which point about 95 percent of the marginal utility estimates lie within 25 percent of the mean.

Most of the parameter estimates behave similarly at a restricted sample size of 120, with around 95 percent of the marginal utility estimates for travel time, on-time arrival, fuel cost, free flow time multiplied by trip distance, and distance-based charges between 20 percent and 30 percent of their respective means. The mean estimate for distance-based charges and the estimated standard deviation of the freight rate (not a design attribute) do show larger divergence from the mean, however, at a restricted sample size of 120 (whilst performing reasonably well at a sample size of 130). Considered together with Table 6, this evidence suggests that a sample size of 130 should have been sufficient to yield significant parameter estimates that are also reliable, but that further reductions in sample size could result in small sample size effects of heterogeneity (i.e., unrepresentative samples) leading to either insignificant parameter estimates or significant parameter estimates that are considerably different from the values that one would expect under a larger sample.

4.2 Results for shippers

Turning to the shipper portion of the study, our analysis centres on a relatively complex base model. In this model, the marginal utility of each attribute in the design (other than waiting time, which is insignificant just as in the transporter model), along with the reference alternative and freight rate, is represented as a random parameter. This flexibility in model selection is likely due in large part to the near doubling of choice observations relative to the transporter model (138 respondents with 8 choice observations each, versus 145 transporters with 4 choice observations each). Such a hypothesis leads directly to the central theme of this section, the potential degree to which the sample could have been restricted whilst still yielding statistically significant parameter estimates for shippers.

We begin the discussion with a look into the discrepancies between the assumed prior parameter values and the observed parameter estimate values in the full sample. Table 8 highlights the differences between the priors and the observed marginal utility parameters:

Table 8: Shipper base model versus priors

(t-statistics in parentheses; all random parameters distributed normally)

	Full Sample	**Marginal Rates of Substitution* (MRS)**	**Prior MRS***
Parameter			
Reference Alternative	0.985 (4.302)		
Free-Flow Time (min) - Mean	-0.016 (-4.095)	1.6	0.828
Free-Flow Time (min) – Std. Dev.	0.020 (4.187)		
Slowed-Down Time (min) - Mean	-0.030 (-3.392)	3.0	0.828
Slowed-Down Time (min) – Std. Dev.	0.026 (1.840)		
Probability of On-Time Arrival (%) - Mean	0.179 (5.549)	-17.9	-2
Probability of On-Time Arrival (%) – Std. Dev.	0.167 (5.126)		
Fuel Cost ($) - Mean	-0.010 (-2.866)	1	1
Fuel Cost ($) – Std. Dev.	0.006 (2.041)		
Freight Rate ($) - Mean	-0.006 (-1.727)	0.6	--
Freight Rate ($) – Std. Dev.	0.020 (4.955)		
Distance-Based Charges ($) - Mean	-0.013 (-3.149)	1.3	2
Distance-Based Charges ($) – Std. Dev.	0.010 (3.191)		
Waiting Time (min)	0 (Not in model)	0	0.828

***-Marginal rates of substitution are for mean values with respect to the mean value for fuel cost.**

The relationships amongst parameter estimates for shippers are generally, and in some cases drastically, different to the assumed prior relationships. The mean marginal rates of substitution with respect to fuel cost reveal much stronger sensitivities to travel time components than assumed, with values of 1.6 and 3.0 dollars per minute in free-flow and slowed-down conditions, respectively, versus prior values of 0.83 for both travel time components. That is, the base model implies that shippers, on average, are both more sensitive to travel time relative to fuel cost than assumed, and more sensitive to slowed-down time relative to free-flow time than assumed. Consistent with the transporter model, shippers appeared to be less sensitive to distance-based charges than assumed, on average, with a mean marginal rate of substitution between distance-based charges and fuel cost of 1.3 dollars spent on fuel per dollar spent on distance-based charges, compared to an assumed rate of 2. Most strikingly, the estimated mean marginal rate of substitution between the probability of on-time arrival and fuel cost is almost eight times higher than assumed, at –17.9 dollars per percentage point in reliability compared to –2. Hence, the assumed prior sensitivity to reliability was much lower than observed.

As with the transporter model, we estimated 100 sub-samples for decreasing multiples of ten respondents starting from the full sample size. The mean parameter estimates under each cohort size are shown in Table 9:

Table 9: Mean bootstrap parameter estimates by cohort size

(shippers, 100 Sub-samples per cohort size, 8 choice observations per respondent)

	All (138)	**130**	**120**	**110**	**100**	**90**	**80**	**70**	**60**	**50**	**40**
Reference Alternative	0.9847	0.9947	0.9959	0.9868	0.9853	0.9704	0.9526	0.9634	0.9392	0.9579	0.9249
Free-Flow Time (min) - Mean	-0.0163	-0.0151	-0.0151	-0.0155	-0.0158	-0.0157	-0.0160	-0.0160	-0.0161	-0.0163	-0.0177
Free-Flow Time (min) – Std. Dev.	0.0203	0.0204	0.0206	0.0209	0.0215	0.0213	0.0225	0.0226	0.0225	0.0226	0.0247
Slowed-Down Time (min) - Mean	-0.0295	-0.0287	-0.0284	-0.0287	-0.0291	-0.0294	-0.0302	-0.0305	-0.0326	-0.0325	-0.0357
Slowed-Down Time (min) – Std. Dev.	0.0256	0.0242	0.0241	0.0246	0.0252	0.0248	0.0270	0.0274	0.0284	0.0288	0.0324
Probability of On-Time Arrival (%) - Mean	0.1785	0.1699	0.1698	0.1739	0.1777	0.1772	0.1795	0.1800	0.1773	0.1802	0.1947
Probability of On-Time Arrival (%) – Std. Dev.	0.1670	0.1627	0.1648	0.1680	0.1714	0.1714	0.1744	0.1754	0.1781	0.1774	0.1935
Fuel Cost ($) - Mean	-0.0100	-0.0082	-0.0085	-0.0089	-0.0096	-0.0089	-0.0097	-0.0105	-0.0113	-0.0120	-0.0119
Fuel Cost ($) – Std. Dev.	0.0065	0.0038	0.0038	0.0043	0.0044	0.0047	0.0049	0.0054	0.0062	0.0066	0.0076
Freight Rate ($) - Mean	-0.0058	-0.0111	-0.0111	-0.0115	-0.0105	-0.0121	-0.0125	-0.0121	-0.0113	-0.0106	-0.0144
Freight Rate ($) – Std. Dev.	0.0199	0.0205	0.0201	0.0210	0.0205	0.0206	0.0209	0.0214	0.0188	0.0211	0.0221
Distance-Based Charges ($) - Mean	-0.0132	-0.0095	-0.0097	-0.0107	-0.0116	-0.0111	-0.0121	-0.0127	-0.0149	-0.0154	-0.0182
Distance-Based Charges ($) – Std. Dev.	0.0102	0.0096	0.0102	0.0099	0.0109	0.0102	0.0111	0.0118	0.0138	0.0147	0.0174

Unlike the transporter model, which demonstrated fairly steady mean parameter estimates across cohort sizes up to around one-half of the full sample, we observe two separate tendencies in the shipper model. The first tendency is the aforementioned tendency, in which approximately half of the attributes’ marginal utility estimates show no major trend as sample size is reduced until the sample size approaches one half of the sample. That is, marginal utility estimates for the reference alternative, free-flow mean effect, slowed-down time mean effect, on-time arrival mean effect, freight rate and distance-based charge standard deviation are generally steady as the cohort size falls from 138 toward around 80.

However, this is countered by the second tendency, in which the remaining explanatory variables tend to reveal a mean estimate that changes quickly from the full sample value as the cohort size falls. This indicates a strong statistical influence by a relatively small group of observations (i.e., outlier effects) over these variables. Indeed, the main empirical models arising from the original study removed some observations to control for outlier effects (chiefly unusually long trips and vehicles with unusual stated fuel economy). The new values taken by the mean parameter estimates at a sample size of 130 appear to represent a steady baseline for these variables, as the marginal utility estimates for these variables demonstrate stability from a sample size of 130 down to around the same threshold as the other attributes. This confirms the general trend throughout both the transporter and shipper sides of the sample for sub-samples to yield unbiased estimates relative to the full sample until the cohort size falls to around 80; in the case of this subset of attributes, this tendency is only apparent once controlling for outlier effects.

We now turn to Table 10 to examine the frequencies with which the marginal utility parameter estimates reached statistical significance at each restricted sample size:

Table 10: Frequency of parameter estimates significant at target confidence levels by cohort size (shippers)

Frequency of Significance at the 95% Confidence Level

	Cohort Size									
Attribute	**130**	**120**	**110**	**100**	**90**	**80**	**70**	**60**	**50**	**40**
Reference Alternative	100	100	100	99	99	96	89	75	67	63
Free-Flow Time (min) - Mean	100	100	100	100	99	98	99	86	79	62
Free-Flow Time (min) – Std. Dev.	100	100	99	99	100	98	97	85	79	57
Slowed-Down Time (min) - Mean	100	100	100	100	95	88	81	72	53	41
Slowed-Down Time (min) – Std. Dev.	18	18	13	11	18	19	19	14	14	17
Probability of On-Time Arrival (%) - Mean	100	100	99	99	100	100	100	100	100	100
Probability of On-Time Arrival (%) – Std. Dev.	100	100	99	99	100	100	100	99	100	91
Fuel Cost ($) - Mean	97	90	82	74	61	53	48	34	32	25
Fuel Cost ($) – Std. Dev.	15	7	18	15	17	16	18	22	20	21
Freight Rate ($) - Mean	87	81	75	59	64	59	58	51	38	41
Freight Rate ($) – Std. Dev.	100	100	99	97	98	99	94	86	84	78
Distance-Based Charges ($) - Mean	98	95	85	75	63	59	52	44	39	34
Distance-Based Charges ($) – Std. Dev.	93	91	78	76	66	57	51	55	40	39

Frequency of Significance at the 90% Confidence Level

	Cohort Size									
Attribute	**130**	**120**	**110**	**100**	**90**	**80**	**70**	**60**	**50**	**40**
Reference Alternative	100	100	100	100	99	99	94	85	81	71
Free-Flow Time (min) - Mean	100	100	100	100	100	100	100	97	92	84
Free-Flow Time (min) – Std. Dev.	100	100	99	99	100	98	100	94	89	78
Slowed-Down Time (min) - Mean	100	100	100	100	98	98	93	86	77	60
Slowed-Down Time (min) – Std. Dev.	59	43	48	35	34	35	30	29	25	25
Probability of On-Time Arrival (%) - Mean	100	100	99	99	100	100	100	100	100	100
Probability of On-Time Arrival (%) – Std. Dev.	100	100	99	99	100	100	100	99	100	99
Fuel Cost ($) - Mean	100	97	88	84	78	72	65	52	47	38
Fuel Cost ($) – Std. Dev.	21	12	25	22	26	25	27	30	26	26
Freight Rate ($) - Mean	92	90	80	71	73	65	62	56	46	55
Freight Rate ($) – Std. Dev.	100	100	99	97	98	99	95	90	91	84
Distance-Based Charges ($) - Mean	100	98	94	90	88	80	73	63	53	43
Distance-Based Charges ($) – Std. Dev.	97	94	86	85	78	69	59	66	56	51

Many of the attributes in the design maintain significance up to and even beyond the 80 respondent level. The mean effects for free-flow time, slowed-down time and on-time arrival probability, along with the standard deviations for free-flow time and on-time arrival probability, each maintain a strong tendency to reach statistical significance at restricted sample sizes. The reference alternative also enters the model significantly in almost all cases down to a sample size of 80. Two heterogeneous effects that are significant in the full model are seldom significant at any of the restricted sample sizes; preferences for slowed-down time and fuel cost appear to be homogeneous at small sample sizes, in general. Indeed, the statistical significance of fuel cost is not assured when the sample size is restricted to around 100. This is not a terribly difficult outcome to justify (i.e., after controlling for the freight rate, the shipper may not be terribly sensitive to fuel cost). However, if the large sample estimates showing a sensitivity to fuel cost are correct, this uncertainty in significance for fuel cost at moderate sample sizes is a concern. Some of this could feasibly be mitigated by re-specifying to marginal utility parameter as non-random, which may increase the explanatory power of the mean effect. The freight rate (not a design attribute) demonstrates significant preference heterogeneity at sample sizes of 70 and even lower. Lastly, the mean marginal disutility of distance-based charges tends to enter the model with a p-value less than 0.1 in sample sizes as low as 90, with preference heterogeneity apparent at a somewhat smaller frequency. This shares the implications for fuel cost, in that whilst a low sensitivity to distance-based charges by shippers may be intuitive after accounting for the freight rate, at larger sample sizes shippers' sensitivities to distance-based charges are indeed statistically significant. Hence, if the model fails to show this under a restricted sample size, it may be a case of observing too few choices to identify shippers' true behaviour.

Turning to Table 11, we observe a rather large spread of most parameter estimates relative to their mean values at even moderately restricted sample sizes:

Table 11: Standard deviation of parameter estimates as a percentage of the mean by cohort size (shippers)

	130	**120**	**110**	**100**	**90**	**80**	**70**	**60**	**50**	**40**
Reference Alternative	0.066	0.111	0.142	0.161	0.178	0.212	0.251	0.342	0.338	0.474
Free-Flow Time (min) - Mean	-0.082	-0.119	-0.137	-0.184	-0.202	-0.237	-0.240	-0.308	-0.368	-0.576
Free-Flow Time (min) – Std. Dev.	0.106	0.143	0.170	0.200	0.227	0.274	0.302	0.351	0.358	0.573
Slowed-Down Time (min) - Mean	-0.074	-0.098	-0.139	-0.169	-0.208	-0.250	-0.284	-0.307	-0.358	-0.479
Slowed-Down Time (min) – Std. Dev.	0.135	0.168	0.222	0.245	0.380	0.381	0.453	0.487	0.561	0.673
Probability of On-Time Arrival (%) - Mean	0.070	0.083	0.109	0.130	0.164	0.198	0.206	0.208	0.280	0.378
Probability of On-Time Arrival (%) – Std. Dev.	0.052	0.071	0.090	0.104	0.133	0.171	0.203	0.218	0.238	0.378
Fuel Cost ($) - Mean	-0.209	-0.292	-0.366	-0.416	-0.508	-0.564	-0.595	-0.736	-0.714	-1.104
Fuel Cost ($) – Std. Dev.	0.542	0.559	0.716	0.760	0.843	0.914	0.963	1.005	1.140	1.187
Freight Rate ($) - Mean	-0.296	-0.346	-0.401	-0.538	-0.543	-0.600	-0.708	-0.942	-1.130	-1.155
Freight Rate ($) – Std. Dev.	0.099	0.142	0.183	0.283	0.303	0.390	0.422	0.497	0.519	0.695
Distance-Based Charges ($) - Mean	-0.263	-0.332	-0.436	-0.469	-0.516	-0.536	-0.598	-0.803	-0.707	-0.970
Distance-Based Charges ($) – Std. Dev.	0.268	0.273	0.416	0.476	0.582	0.622	0.698	0.846	0.867	0.997

This is important, in that the preceding tables show strong tendencies for the marginal utility estimates to reach statistical significance even when the sample size is reduced to around 80 respondents. Hence, under a restricted sample there would have been a risk of observing significant parameter estimates that would be considerably different from the values found under a large sample. The lowest sample size that yielded a reasonably tight set of parameter value distributions was 110; at this sample size, around 95 percent of the sub-samples taken resulted in parameter estimates with 36 percent of the mean for the reference alternative, free-flow time (both mean and standard deviation), standard deviation of slowed-down time, on-time arrival probability (both mean and standard deviation), and the standard deviation of the freight rate. The distributions for the standard deviations of free-flow time and the freight rate become wide at a sample size of 100, with the distributions of the mean estimates for free-flow and slowed-down time growing in spread at a sample size of 90. At a sample size of 80, the distributions for all but on-time arrival probability (mean and standard deviation) are wide enough that a spread of 40 percent of the mean is insufficient to account for 95 percent of estimates.

Ultimately, it appears that the design itself was strong enough to accommodate a sample size as small as 80 shippers; the choice observations that would have been captured under a sample of 80 shippers would likely have led to an econometric model of shipper choice behaviour that yielded statistically significant parameter estimates. However, the strength of the design would not necessarily have been sufficient to obtain marginal utility estimates that are reasonably close to the values that would be obtained under a larger sample. This may be less an issue of optimal design itself, and more an issue of heterogeneity dominating statistical efficiency concerns under small sample sizes.

5. Conclusions

The experimental design for the freight study was viewed as a strong success upon the completion of the survey, in that it led to the capture of sufficient preference information to estimate a series of complex econometric models despite the limited sample size of 145 transporters (with 580 choice observations) and 138 shippers (with 1106 choice observations). The bootstrapping exercise examined in this paper served to gauge just how far the design could have been pushed, had there been greater difficulty in sourcing respondents for the study. The initial analysis indicates that the design would likely have been sufficient to yield accurate and significant behavioural implications had the sample been restricted to around 130 transporters and 100 shippers. If the sample had been restricted to between 80 and 100 respondents from both decision-making classes, the sample size may have been small enough to yield an unrepresentative sample whilst still offering enough choice observations under the optimal design to achieve statistically significant parameter estimates. In such an unfortunate case, the significant parameter estimates could have been biased away from the values that would be found under a larger sample.

Hence, it is important to acknowledge that, whilst optimal designs can be a powerful tool in achieving statistically significant parameter estimates under small sample sizes, behavioural factors can outweigh statistical factors in determining an appropriate sample size. Ultimately, although statistical significance is a necessary condition for identifying preference information, it may not be sufficient. Rather, one must ensure that stability in parameter estimates has been reached before one can have confidence that the statistically significant parameter estimates obtained are also plausible estimates.

We are confident that we have found such stability in the estimates obtained in the study at levels near the full sample and even restricted as low as around 100 respondents of both classes of decision makers. Some instability found in interaction terms and random parameter distributions could be remedied through alternative model specifications, further solidifying the models under restricted sample sizes.

In addition to the implications found relating to each half of the sample, the bootstrapping exercise revealed some interesting implications regarding experimental designs for studies involving multiple classes of decision makers, in general. In the freight study, the restriction of having capturing half as many choice observations per transporter compared to each shipper led to the design algorithm sacrificing some statistical efficiency with respect to the choices made by shippers over the choice sets faced jointly by transporters and shippers. This allowed the experimental design to have a greater ability to ensure statistical significance for transporters under a smaller number of choice observations relative to a design that weighted the prior information on transporters and shippers equally. The design appears to have successfully struck such a balance, in that both the transporter and shipper models demonstrated similar rates of decline in performance as the sample size was decreased. That is, despite the fact that the transporter model was calibrated against one-half the number of observations for a given sample size relative to the shipper model, statistical confidence in the behavioural implications for both sides of the sample decreased in a similar manner as the sample size was increasingly restricted.

This similarity across the two decision-making classes likely reflects one of two effects. The first effect would be the optimal one, in which the design accomplished what it was intended to accomplish. If this is true, we would recommend the same approach in similar studies involving multiple decision-making classes. A caveat should be issued, however, in that it may be appropriate to find an additional weighting mechanism in the experimental design process if one would expect to have greater difficulty in sourcing respondents from a particular class of decision maker. In such a case, rather than using the number of choice observations per survey instrument for each class as a primary weighting criterion, the expected ratio of choice observations to be obtained in the study across classes may be important to consider. For example, in the freight study, if one had expected to only have the ability to recruit half as many shippers as transporters (and hence obtain an equal number of choice observations for the two classes), the design would likely have been improved if it had been calibrated to weight each class equally.

The other effect that may have resulted in similar behaviour across the two models as the sample size was increasingly restricted is that the shipper model is more complex than the transporter model, and hence requires more data than the transporter model to reach a desired level of statistical significance. This is certainly a plausible explanation for at least some of the similarity in performance across the models, and it would be beneficial to re-examine the design under a simpler, common modelling structure for both transporters and shippers. We selected the more complex shipper model in this study because it takes the same form as in the original freight study, and because it allows us to examine a more complex model under restricted sample sizes. We will isolate this effect in ongoing research by establishing an appropriate common modelling structure and conducting repeated bootstrapping exercises as in this study.

In other ongoing research, we will attempt to account for flexibility in model specification as sample size is reduced, giving a more thorough picture of the potential for an optimal design to produce meaningful inference under small sample sizes. We will also test the performance of a range of orthogonal designs as a benchmark to reveal the sample sizes that would be required to achieve the same quality of inference in the absence of an optimal design. These exercises should demonstrate further the value and limitations of optimal designs in econometric studies of choice behaviour.

References

Alpizar, F., Carlsson, F., and Martinsson, P. (2003). Using choice experiments for non-market valuation, *Economic Issues*, vol. 8, 83-110.

Bliemer, M.C. and Rose, J.M. (2006). Designing stated choice experiments: The state of the art, accepted for presentation at 11th International Conference on Travel Behaviour Research - Kyoto, August 16-20, Japan.

Bliemer, M.C. and Rose J.M. (2009). Efficiency and sample size requirements for stated choice experiments, *The Transportation Research Board (TRB) 88th Annual Meeting*, Washington, D.C., United States, 15th January 2009.

Bliemer, M.C. and Rose, J.M. (2008). Construction of experimental designs for mixed logit models allowing for correlation across choice observations, *Proceedings of the 87th Annual Meeting of the Transportation Research Board*, Washington D.C., January.

Brewer, A. and Hensher, D. A. (2000). Distributed work and travel behaviour: the dynamics of interactive agency choices between employers and employees, *Transportation*, 27 (1), 117-148.

Bunch, D.S., J.J. Louviere and D.A. Anderson (1994). A comparison of experimental design strategies for multinomial logit models: The case of generic attributes, working paper, Graduate School of Management, University of California at Davis.

Carlsson, F. and Martinsson, P. (2003). Design techniques for stated preference methods in health economics, *Health Economics*, vol. 12, no. 4, 281-294.

Hensher, D.A. and Puckett, S.M. (2007). Power, concession and agreement in freight distribution chains subject to distance-based user charges, *International Journal of Logistics: Research and Applications* vol. 11, no. 2, 81-100.

Huber, J., and K. Zwerina (1996). The importance of utility balance and efficient choice designs, *Journal of Marketing Research*, 33, 1996, 307-317.

Kanninen, B.J. (2002). Optimal design for multinomial choice experiments, *Journal of Marketing Research*, 39(2), 214-217.

Kuhfeld, W., Tobias, R. and Garrat, M. (1994). Efficient experimental design with marketing research applications, *Journal of Marketing Research*, vol. 31, 545-557.

Manski, C. F. (2000). Economic analysis of social interactions, *Journal of Economic Perspectives*, 14(3), 115-136.

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

McFadden, D. (2001a). Economic choices, *American Economic Review*, 91 (3), 351-378.

McFadden, D. (2001b). Overview of the Invitational Choice Symposium, Asilomar Conference Center, California, June.

Puckett, S.M. and Hensher, D.A. (2008). The role of attribute processing strategies in estimating the preferences of road freight stakeholders, *Transportation Research Part E*, vol. 44, 379-395.

Puckett, S.M., Hensher, D.A., Collins A. and Rose, J. (2007). Design and development of a stated choice experiment in a two-agent setting: interactions between buyers and sellers of urban freight distribution services, *Transportation*, vol. 34, no. 4, 429-451.

Sándor, Z., and M. Wedel (2001). Designing conjoint choice experiments using managers' prior beliefs, *Journal of Marketing Research*, 38, 430-444.

Appendix

A1: Choice sets common across sampled group members – trips of two hours or less (attribute levels relative to reference alternative)

Choice Set	Alternative	Free-Flow Time	Slowed-Down Time	Waiting Time	Fuel Cost	Likelihood of On-Time Arrival	Distance-Based Charges
1	B	-25%	+50%	-25%	-25%	0	-50%
1	C	0	-50%	+25%	0	+50%	-25%
2	B	-25%	-25%	0	-50%	0	+25%
2	C	+50%	+50%	+25%	+50%	+25%	-50%
3	B	0	-50%	+25%	+25%	+50%	+50%
3	C	-25%	-25%	0	0	0	-25%
4	B	+25%	+50%	+50%	+25%	0	0
4	C	+50%	-25%	0	+25%	+50%	-25%

A2: Choice sets common across sampled group members – trips of more than two hours (attribute levels relative to reference alternative)

Choice Set	Alternative	Free-Flow Time	Slowed-Down Time	Waiting Time	Fuel Cost	Likelihood of On-Time Arrival	Distance-Based Charges
1	B	-25%	-25%	-25%	+50%	+50%	-50%
1	C	-25%	0	+50%	-50%	+50%	-50%
2	B	0	-25%	+25%	0	0	+50%
2	C	-25%	+50%	-25%	+50%	+50%	-50%
3	B	-50%	0	+25%	+50%	-50%	-50%
3	C	-50%	+50%	-50%	-25%	0	+50%
4	B	0	-50%	-25%	+50%	+50%	+50%
4	C	+25%	+25%	+25%	+50%	+25%	+25%

A3: Choice sets for shippers only – trips of two hours or less (attribute levels relative to reference alternative)

Choice Set	Alternative	Free-Flow Time	Slowed-Down Time	Waiting Time	Fuel Cost	Likelihood of On-Time Arrival	Distance-Based Charges
5	B	+25%	0	-50%	-50%	+25%	0
5	C	-50%	+25%	+50%	+50%	-50%	+50%
6	B	+50%	+25%	-50%	-25%	+25%	-25%
6	C	-50%	-25%	-25%	+25%	-50%	+25%
7	B	-50%	+50%	+50%	0	+25%	0
7	C	-25%	-25%	0	-25%	-25%	-50%
8	B	-50%	+25%	0	+25%	-25%	0
8	C	+25%	-50%	-25%	-50%	0	-25%

A4: Choice sets for shippers only – trips of more than two hours (attribute levels relative to reference alternative)

Choice Set	Alternative	Free-Flow Time	Slowed-Down Time	Waiting Time	Fuel Cost	Likelihood of On-Time Arrival	Distance-Based Charges
5	B	+25%	0	-50%	+50%	-50%	+50%
5	C	+25%	-25%	-25%	-25%	+50%	+25%
6	B	+50%	+25%	+25%	0	0	-25%
6	C	-25%	+50%	-50%	+25%	+25%	0
7	B	-50%	-50%	+50%	+25%	+50%	+25%
7	C	+25%	0	+25%	-50%	-25%	0
8	B	0	-50%	-50%	+25%	-50%	-25%
8	C	-50%	0	-50%	+25%	+25%	-25%