Chapter 10

Experimental design and the estimation of willingness to pay in choice experiments for health policy evaluation

Richard T. Carson and Jordan J. Louviere

10.1 **Introduction**

This chapter focuses on stated preferences obtained from discrete choice experiments (DCEs also known as SPDCEs), as opposed to data that reflect real market choices (revealed preferences (RP) as discussed in Chapter 9). DCEs try to simulate the essential elements of real market options that consumers might face in the future. Unlike real market choice data, DCEs rely on constructed markets in which key factors that are hypothesized to drive choices are systematically varied. To the extent that the consumers in a DCE make choices in a manner consistent with the way in which they would actually choose in a real market, one can derive standard welfare estimates for policy changes. The remainder of this chapter is devoted to discussing and illustrating how this can be accomplished with DCEs. More details on DCEs can be found in Louviere, Hensher, and Swait (1).

In order to collect DCE data information from consumers, one must identify factors that drive the choices of interest. These factors are called 'attributes' of the choice options. Once these attributes are identified, one must assign them values, known in experimental design parlance as 'levels'. Taken together, the attributes and levels define and determine possible choice options that can be offered to consumers in a DCE survey. That is, a factorial combination of attribute levels completely defines the possible choice options. So, for example, if there are three attributes, say A(4), B(3), and C(2), with the associated number of levels in parentheses, the factorial combinations, or all possible options are $4 \times 3 \times 2 = 24$. We refer to that factorial combination as a 'full' or 'complete' factorial design. Typically, the number of combinations in a full factorial design is too many to use in practical field applications of DCE surveys, so one has to sample from the full factorial to reduce the size of the problem. There are many ways to sample from full factorials, but a common approach in DCE surveys is to sample based on what is known as a 'fractional factorial design'. We return to these ideas in more detail later in the chapter.

The type of experimental design used to construct a DCE survey is important for three reasons: 1. it determines the economic quantities of potential interest that can be statistically identified from an estimation perspective; 2. it strongly influences confidence intervals associated with these quantities (statistics) given a fixed sample size or equivalently, influences the sample size needed to achieve a given confidence interval or level of precision for a statistic, and so plays a major role in the cost of a project; 3. the attributes/levels can influence the plausibility of questions; and hence, influence the quality of data obtained. The first two issues are largely statistical in nature and are the subject of this chapter.[1] The third issue deals more with issues of survey design and are not considered further here, although it may place constraints on what can be measured in a DCE survey.[2]

This chapter uses examples from health policy research, but these examples also have application in analyzing consumer choices in areas like culture, environment, transport, utilities, or more generally where government has a strong role in determining which options are available and/or how they are priced, like health. Overviews of the use of DCE methods to value health issues can be found in various sources (e.g., 8–11).[3]

We begin by considering a single binary choice (SBC) contingent valuation question because many basic issues associated with experimental design can be easily seen in the context of the SBC format (see, e.g., (13)). This question format is popular in the stated preference literature and enjoys a number of desirable incentive properties under certain conditions.[4] This question format also is the simplest example of a more general discrete choice experiment format (see, e.g., (1)). The key properties of the SBC format from our perspective are:

1. Only one choice question [set] is asked;
2. The question asked offers only two alternatives; and,
3. Only one attribute of the scenario, typically cost, is varied across respondents.

For example, consider the following simple proposition. A large university in the USA is considering offering employees the option of purchasing a dental plan. The dental plan covers 80% of normal costs associated with all standard, non-cosmetic dental procedures. The university is interested in what fraction of employees will subscribe to the dental plan at various prices. A DCE for this policy problem would describe the

[1] The two key statistical properties of an experimental design are identification and precision. Louviere, Hensher, and Swait (1) note two other properties that can influence the desirability of a design that are not statistical in nature. These are cognitive complexity and market realism. Both considerations can restrict the nature of the attributes and design used.

[2] There are several general books on survey design (e.g., 2–4) but there is a surprising lack of guidance on the issue of SP choice questions in a policy context. Some exceptions are (5;6). A small but growing literature (e.g., (7) looks at the implications of task complexity in choice experiments.

[3] For an overview of different health valuation methods including SP methods with an emphasis on determining the value of a statistical life, see (12).

[4] The SBC question format was recommended by the Arrow *et al.* (14) panel that looked at SP methods for valuing natural resource damages for the U.S. Government. Carson and Groves (15) examined the incentive properties of different types of SP questions in detail.

dental plan in sufficient detail for employees to understand what they would be purchasing, provides the monthly cost associated with the plan and asks the question 'Would you subscribe to this plan if offered to you?'. The employee can choose to accept the plan or reject it (the two options). In this case, a single attribute is varied over a range of levels, namely the cost attribute. This allows the analyst to trace out the fraction of employees who would subscribe at each presented level of cost.

In more general DCE formats, one often sees: (a) more than one choice set asked, (b) more than two alternatives offered, and (c) more than one attribute varied. Frequently, however, only one or two of the three generalizations of the SBC format are used, so it is useful to keep in mind the specific generalization of the SBC format when thinking about issues involving DCEs. The nature and properties of different experimental designs become more important as DCE formats grow more complicated and we systematically illustrate where new issues arise and how the statistical models that can be estimated are tied to the experimental design used to construct a SP survey.[5]

From an applications standpoint, we focus on three cases:

1. A new good may be provided, and if it is provided, the person using it has to pay for it (everyone pays if it is a pure public good provided via coercive taxation). Interest lies in estimating total willingness to pay (WTP) for having the good supplied rather than the current status quo good.
2. One wishes to estimate the WTP to have one or more new alternatives added to a set of choices available to a consumer.
3. One wishes to estimate the WTP for a change in one or more of the attributes of an alternative.

We begin our discussion by laying out the theoretical welfare measures for the three above cases. We then introduce the basic concepts of experimental design in the context of the SBC format. Next, we illustrate the issues involved in moving to different types of DCE formats. Finally, we try to provide guidance to applied researchers who want to conduct a DCE study using reasonably efficient designs where the statistics of primary interest are statistically identified.

10.2 Economic welfare measures for health policy changes

SP surveys analyzed from a random utility perspective often aim to produce estimates for policy purposes; hence, we briefly review the theory relating to welfare economic measures of value. Literature on this topic is vast, so to more fully appreciate the issues involved, interested readers may wish to pursue comprehensive treatments in (16–18).

We begin by denoting the item being valued by q and initially treat this as a single item that could be a commodity or a program involving some mix of commodities

[5] There are many other relevant cases for measuring changes in welfare. In particular, there are a set of analogous measures that focus on minimum willingness to accept (WTA) compensation for undesirable changes (see Chapter 6 & 7).

treated as a fixed group – the key feature is that q is a scalar. Later, we will let q consist of a bundle of attributes and we will ask questions about how a change in one attribute influences economic values. The latter does not change the underlying framework as one can define two distinct q's that differ only in a change in one attribute of interest. We assume that a consumer has a utility function defined over quantities of various market commodities, denoted by the vector x, from which a consumer can freely choose and the item q. Thus, the direct utility function is given by $u(x, q)$. Often analysts work with the corresponding indirect utility function, $v(p, q, y)$, where p is the vector of the prices of the market commodities x, and y is the consumer's income.[6] We make the conventional assumption that $u(x, q)$ is increasing and quasi-concave in x, which implies that $v(p, q, y)$ satisfies the standard properties with respect to p and y[7]; but we make no assumptions about q. If the agent regards q as a 'good', $u(x, q)$ and $v(p, q, y)$ both will be increasing in q; if it is regarded as a 'bad', $u(x, q)$ and $v(p, q, y)$ both will be decreasing in q; and if the agent is indifferent to q, $u(x, q)$ and $v(p, q ,y)$ both will be independent of q. We make no assumption about quasi-concavity with respect to q.

The act of valuation implies a contrast between two situations: one with item q, and one without q. This is an important concept because economic valuation always involves a comparison/tradeoff between two or more situations where the 'or more' part always can be rewritten as a set of binary comparisons. We interpret what is being valued as a change in q.[8] Specifically, suppose that q changes from q^0 to q^1; the consumer's utility changes from $u^0 \equiv v(p,q^0,y)$ to $u^1 \equiv v(p,q^1,y)$. If she sees this change as positive, $u^1 > u^0$; if she sees it as negative, $u^1 < u^0$; and if she is indifferent, $u^1 = u^0$. The value of the change to her in monetary terms is represented by the Hicksian income compensation measure, C, which is the amount of money that satisfies:

$$v(p,q^1,y-C) = v(p,q^0,y)\,.^9 \tag{10.1}$$

---

6 The definition of income is always problematic in empirical work. Ideally, it refers to some notion of permanent household supernumerary income, which is disposal permanent income less total expenditure on subsistence minima and previously committed expenditures.

7 That is, we assume $v(p, q, y)$ is homogeneous of degree zero in p and y, increasing in y, non-increasing in p, and quasiconvex in p.

8 The alternative is to represent it as a change in p. McConnell (19) adopts this approach for a valuation question of the form 'Would you accept a payment of \$A to give up your right to use this commodity for 1 year?' Let p^* be the choke price vector (*i.e.*, a cost vector such that, at these costs, the individual would choose not to consume the resource), and let p^0 be the baseline price vector. McConnell represents the change as a shift from (p^0, q, y) to (p^*, q, y).

9 Note that we also can get the Hicksian equivalence measure, which in this case is WTA for giving up the right to q^1. If the sign of $u^1 - u^0$ is positive then WTP and WTA will both be positive. The Hicksian income compensation function often is formally defined as the difference between two (Hicksian) expenditure functions, another alternative representation of the direct utility function that describes how much income is needed to achieve a specified level of utility given a price vector for marketed goods and the level of q. $C = m(p, q^0, u^0) - m(p, q^1, u^0)$. The first term on the right is simply equal to y.

To emphasize the dependence of the compensating measure on (i) the starting value of q, (ii) the terminal value of q, and (iii) the value of (p, y) at which the change in q occurs, we sometimes write it as:

$$C = C(q^0, q^1, p, y)\,. \tag{10.2}$$

For a desirable change which the consumer does not have the property right to enjoy without paying

$$\text{WTP} = C(q^0, q^1, p, y)\,. \tag{10.3}$$

One can parameterize either the WTP function directly (20) or begin with a parameterization of the underlying utility function (21). Typically, applied researchers fit simple linear or logarithmic functions, but one can fit more complex utility functions, and SP data often is better suited for this than RP data. We illustrate these concepts with a specific example, the Box–Cox indirect utility function:[10]

$$v_q = \alpha_q + \beta_q\left(\frac{y^\lambda - 1}{\lambda}\right), \quad q = 0,1\,, \tag{10.4a}$$

where $\alpha_1 \geq \alpha_0$ and $\beta_1 \geq \beta_0$. Equation (10.4a) can be regarded as a form of CES utility function in q and y with λ being the income elasticity of WTP. The corresponding formula for C is

$$C = \left(\frac{\beta_0 y^\lambda}{\beta_1} - \frac{\lambda\alpha}{\beta_1} + \frac{\beta_1 - \beta_0}{\beta_1}\right)^{\frac{1}{\lambda}}, \tag{10.4b}$$

where $\alpha \equiv \alpha_1 - \alpha_0$. McFadden and Leonard (22) employ a restricted version of this model with $\beta_1 = \beta_0 \equiv \beta > 0$, yielding

$$v_q = \alpha_q + \beta\left(\frac{y^\lambda - 1}{\lambda}\right), \quad q = 0,1, \tag{10.5a}$$

$$C = y - \left(y^\lambda - \frac{\alpha}{b}\right)^{\frac{1}{\lambda}}, \tag{10.5b}$$

where $b \equiv \beta/\lambda$. This specification is somewhat flexible as it permits a variety of income elasticities of WTP; the income elasticity of WTP is negative when $\lambda > 1$, zero when $\lambda = 1$, and positive when $\lambda \leq 0$. It also nests many utility models in the existing literature. For example, if $\lambda = 1$, C equals the familiar ratio of $-\alpha/\beta$ often associated with a measure of mean WTP from a logit or probit model.[11]

Now, consider the case where there is more than one possible alternative to the status quo good. As long as a status quo good will remain available and a consumer

[10] For simplicity, we suppress p and write the indirect utility function as a function of q and y; however, α_q and/or β_q are in fact functions of p and z.

[11] Often there are measurement issues with respect to y that play a major role in econometric estimation, pushing empirical researchers to assume that $\lambda = 1$.

will only get utility from at most one good that is an alternative to the status quo good then the economic value of the set of alternatives is simply the maximum WTP defined over all binary paired comparisons involving the status quo good. If the status quo good is chosen in a deterministic setting, WTP equals zero. Otherwise, the economic value equals the maximum amount of money a consumer has to pay for the most preferred of the alternatives relative to being indifferent between choosing the status quo good and the most preferred alternative.[12]

One can make q a function of a bundle of attributes in the sense originally explicated by Lancaster (23). Here α_q is replaced by a function, $g(z)$, where z is a vector of attributes. Typically, $g(z)$ is represented as an additive linear function of the individual attributes some of which are potentially indicator variables, i.e., $\gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3$. In this case, there is nothing particularly special from a welfare economic view from moving from a change in a status quo good versus compared to changes in the attributes of goods other than having to choose the functional form that specifies how the attributes enter the utility function.[13] Other common ways to specify the attribute function are: (a) an additive linear function of the logs of the individual attributes, (b) an additive linear function of the individual attributes plus the first-order interactions between the individual z, and (c) an additive linear representation like a Translog utility function that is a second order approximation to an unknown function that include squares of individual attributes and first order interaction terms between attributes. When there are attributes in the model one often is interested in the marginal effect on WTP of a change in z_i, which is simply $\partial C/\partial z_i$, or the relative effect of a marginal change in one attribute relative to another attribute that can be scaled in a comparable way by looking at $(\partial C/\partial z_i)/(\partial C/\partial z_j)$.

10.3 Going from choice to WTP estimates

SP questions measure a consumer's WTP (or WTA) for change in q or a discrete indicator related to WTP. The utility theoretic model of consumer preference outlined above provides a way to interpret responses to these questions. From a statistical modelling viewpoint, the convention is to treat the survey responses as the realization of a random variable. So, it is necessary to recast the deterministic model of WTP outlined above into stochastic models that can generate probability distributions of the survey responses. Mapping from a deterministic WTP model to a probabilistic model of survey responses involves two steps: 1. adding a stochastic component into

[12] The case where a consumer would use more than one of the alternatives is beyond the scope of this chapter and rarely is examined in SP choice models. Such cases are often dealt with by going to some type of allocation model or bundling alternatives so that bundles are mutually exclusive.

[13] The only 'attribute' that plays a special role is a good's cost. A consumer does not get utility from the cost of a good per se, but instead via the effect of a good's cost on income. In most empirical work this distinction is ignored or an implicit assumption is made about the marginal utility of income so that the only thing that enters the utility function for two (or more) goods is the difference in the cost of the goods.

the deterministic utility model that leads to what is called a *WTP distribution* and 2. specifying a connection between the WTP distribution and what we will call the *survey response probability distribution* based on the assumption of a utility-maximizing response to the survey question. We denote the WTP cumulative distribution function (cdf) as $G_C(z_p)$; for a given individual it specifies the probability that the individual's WTP for the item in question is less than the cost z_p, and we now use the convention of denoting attributes of the good as z_i and the special attribute of cost as z_p:

$$G_c\left(z_p\right) \equiv Pr\left(C \leq z_p\right), \tag{10.6}$$

where the compensating variation, C, is now viewed as a random variable.[14] The corresponding density function is denoted as $g_C(z_p)$.

We illustrate this via a simple example with a *closed-ended, single-bound* discrete choice format. That is, a respondent is asked 'Would you favour a change from q^0 to q^1 if it would cost you z_p?' Suppose the response is 'yes'. This means that the value of C for this individual is some amount more than z_p. In terms of the underlying WTP distribution, the probability of obtaining a 'yes' response is given by

$$Pr\left(\text{Response is 'yes'}\right) = Pr\left(C \geq z_p\right) \equiv 1 - G_c\left(z_p\right). \tag{10.7}$$

Note that a response to this question does not reveal the exact value of C, but instead provides information that C lies in an interval bounded from above or below by z_p.[15]

There are two basic sources of the stochastic component: (a) factors related to the nature of the good or the consumer that influence choice, and are known to the consumer but unknown to the analyst (e.g., 25;26) and (b) a true random component potentially including recording and optimization errors.[16] These two sources effectively are equivalent from the perspective of the simplest statistical estimators, but they have quite different implications for WTP estimates. While source (a) leads to the well-known model of random utility maximization (RUM) in which error components play an integral role in the estimate of summary statistics involving WTP distributions; in contrast, it is desirable to purge source (b) error components from WTP estimates. These two views of error sources also have different implications for what might be observed. Under (a) the probability of picking a dominated alternative should be zero, while under (b) some respondents should pick dominated alternatives with positive probability. For more complex statistical estimators (a) leads one to try

[14] For now, we assume the change is regarded as an improvement so that C measures WTP.

[15] Other relevant information may help to more tightly bound the interval in which the consumer's WTP lies. For example, if the alternative cannot be a 'bad', it may be reasonable to assume a distribution for WTP with no support in the negative range. Carson and Jeon (24) look at ways to use constraints on the upper end related to income.

[16] Excellent discussions of the two perspectives are provided by Hanemann (21) and Cameron (20). McConnell (19) lays out the relationship between the two from the perspective of estimating welfare measures. It is important to note that analysts often claim to estimate a RUM model but use measurement error perspectives when calculating WTP measures.

to allow for heteroscedasticity in preference parameters, while (b) leads one to try to model the error term to allow for heteroscedasticity in some fashion.[17]

The RUM approach proceeds by specifying a particular indirect utility function $v(p,q,y;\varepsilon)$ and a particular distribution for ε. An example of a RUM version of the restricted Box–Cox model is

$$u_q = \alpha_q + \beta\left(\frac{y^{\lambda}-1}{\lambda}\right) + \varepsilon_q \qquad q = 0,1, \tag{10.8a}$$

where ε_0 and ε_1 are random variables with a mean of zero. Consequently,

$$C = y - \left(y^{\lambda} - \frac{\alpha}{b} - \frac{\eta}{b}\right)^{\frac{1}{\lambda}}, \tag{10.8b}$$

where $\alpha \equiv \alpha_1 - \alpha_0$, $b \equiv \beta/\lambda$, and $\eta \equiv \varepsilon_1 - \varepsilon_0$.

In contrast, if we take the second view and operationalize it with an additive error term, we would have

$$C = C(q^0, q^1, p, y) + \varepsilon\,. \tag{10.9}$$

In the case of the Box–Cox model (7), for example,

$$C = y - \left(y^{\lambda} - \frac{\alpha}{b}\right)^{\frac{1}{\lambda}} + \varepsilon\,. \tag{10.10}$$

Comparison of (10.8b) with (10.10) illustrates the difference between the two approaches to formulating a WTP distribution. Inserting an additive random term in utility function (10.8a) leads to a random term that enters into the formula for C in a *non-additive* manner. Even when both approaches lead to similar estimates for mean and median WTP, the implied pdf's may be quite different, particularly in the tails.

Often there can be substantial problems in empirically estimating WTP measures from discrete choice data from either survey choices or market choices.[18] The problems largely stem from the fact that in all discrete choice models the parameters are identified only up to a scale factor.[19] Because of scale, WTP estimates are a ratio of parameter estimates; hence, they can be ill-behaved even if the individual parameter estimates are normally distributed (as suggested by the theoretical foundation

[17] Under (a) the main source of heterogeneity typically is assumed to be differences in preferences, while under (b) the main source of heterogeneity typically is assumed to be differences in ability to answer questions.

[18] Indeed, it usually is better to work with SP data than RP data because the cost variable typically has a much more limited range and there are high correlations between various attributes.

[19] In the case of binary discrete choice, where only cost is varied, one can use non-parametric techniques to avoid some of the problems associated with the scale issue. However, these techniques give much coarser estimates and have not been generalized to the multinomial choice case. See (18) for a discussion.

underlying maximum likelihood estimation) because the ratio of two normal variables is distributed as a Cauchy distribution (although this can be simulated). Further, there is a very tight relationship between the functional form assumed for the cost (z_p) variable and the assumed distribution of WTP.[20] For example, it is common to specify $ln(z_p)$ in a logit model, which implies that the WTP distribution is log-logistic. Unfortunately, this distribution has an (implausible) infinite mean for a wide range of estimated parameter values, although it typically cannot be distinguished from a log-normal (or a Weibull) that has a finite mean in terms of statistical fit.

The problem is that similarly shaped WTP distributions over a wide range of monetary values may have very different behaviour in the far tails,[21] and using more flexible functional forms to allow heterogeneity in preferences can exacerbate problems. For example, if one specifies a cost variable as a random effect and assumes the effect to be normally distributed, it typically implies that some consumers have a negative WTP even if this is implausible. This may concentrate a large fraction of the distribution near zero, causing traditional formulas for mean WTP to blow up.[22]

Often problems in estimating mean WTP are not reported because the analyst assumes them away by estimating a logit or probit model with a linear specification for the cost variable, forcing mean and median WTP to be equal. A similar problem occurs if one estimates a model with the log of cost as a regressor but mistakenly assumes that the correct formula for mean WTP is $\text{EXP}[-\alpha/\beta]$, where α is the constant term (assuming no other attributes) and β is the coefficient on log(cost). This is the correct formula for median WTP but the correct formula for the mean includes a function of the variance, such as the following for a normal distribution: mean WTP = $\text{EXP}[-\alpha/\beta]\text{EXP}[1/2\beta^2]$. Often a better solution to this problem is to recognize that percentiles of the distribution, including the median, usually can be reliably estimated far out in the tails. Traditional welfare economics focuses on mean WTP but policymakers typically care about more than one summary statistic of the WTP distribution.

10.4 **Experimental design for a single binary discrete choice question**

The simplest case for experimental design of a choice experiment occurs when one asks a single binary discrete choice *CV* question of each respondent and only one attribute (typically cost (z_p)) is varied, as earlier noted. Collection of discrete choice

[20] Typically, analysts use the cost of the alternative instead of the more theoretically suitable income minus cost, which is justified by particular assumptions about marginal utility of income. A large amount of measurement error in income also may offset the theoretically desirable properties.

[21] Cost amounts in the far tails are rarely if ever observed in market data and it may be implausible to ask respondents about them in SP surveys.

[22] The key issue is that a non-trivial fraction of consumers may be indifferent to the introduction of any of the alternatives to the status quo, leading to a spike at zero that formally can be modeled as a mixture distribution (27).

data requires the use of a set of design points that represent the cost to agents who are then randomly assigned to those design points. Choice of these design points can greatly influence how many observations are required for a given level of statistical efficiency, which is often referred to as the precision of the estimate.[23]

We begin by considering a linear regression model measuring WTP as a function of changes in a single design factor, say the number of treatments an insurance plan would pay for, z_i, where the line goes through the origin if the value of the design factor is zero. Now, we ask the question 'if you have *n* observations and can run the experiment at two values of the factor, what values would you chose and how many observations should you allocate to each to minimize the confidence interval of the WTP estimate at a particular level of the factor' (i.e., the estimated slope parameter times the factor level of interest)? In this case, the confidence interval for WTP is simply a function of the confidence interval for the slope parameter, so one should choose two values of the factor that are as far apart as feasible. In the case of DCEs, the two values should be chosen to be as far apart as is plausible to respondents. One should allocate half the sample to each of these two values; and it is straightforward to show that this minimizes the confidence interval on the slope parameter.[24] This is a desirable property of a simple DCE because in cases where the expected response to cost is linear, one only needs two levels of cost to accurately estimate the slope. The trick is to ensure that the two points are placed sufficiently far apart to cover much of the response distribution, but not so far into the tails of the distribution that one observes only a few choices.

For example, consider a plan where as before WTP for the plan varies with the number of treatments paid for, but the plan also has other fixed benefits (e.g., information, access to other services at discounts) that do not vary with the number of treatments. Let us represent the WTP for a plan with no treatments by α. Now, WTP for a plan is represented by $\alpha + \beta z_i$, and the objective is to minimize the confidence interval around this quantity for a particular z_i. One can do this by choosing two values for z_i that are as far apart as possible because the confidence interval for α also is minimized by this choice.

Much of this basic intuition extends to binary discrete choice models with a single factor, typically cost. DCEs for these models are analogs of dose–response experiments in medical and related applications; For example, instead of 'cost', experimenters vary the magnitude of a dose of – say – an insecticide, and analytical interest focuses on the percent of the sample population alive falls as dose amount increases. In the case of a DCE for – say – a dental insurance plan, the 'dose' is the levels of cost, and analytical interest focuses on the fraction still in 'favour' as cost increases. Different choice models

23 Like survey design, experimental design is not generally taught in economics departments. A classic text is (28). A more modern, comprehensive reference is (29) or (30).

24 The slope parameter is proportionate to the reciprocal of the square root of $\Sigma_i(z_i - E(z_i))^2$. Given any finite constraint on how far apart the two values of z_i can be from each other, it is possible to show that this quantity is maximized by placing half of the z_i at each end of the constrained distance.

have different likelihood functions and most are non-linear in the model parameters, which has four major implications:

1. The curvature of the likelihood function for most commonly used choice models suggests that the design points should be closer together than in a traditional linear model, but the general principle that they should not be very close remains. The main caveat is not to place the design points too far in the tails because there is too little density to accurately measure the choice probabilities in samples of reasonable size.
2. The optimal design will depend on the number of parameters in the underlying distribution.
3. The optimal design also will depend on the values of those parameters. Generally, one does not need more design points than parameters, but to be able to test more general distributions than one assumes, more design points are needed. Yet, the general principle is that if one fits a parametric distribution characterized by a small number of parameters, one should have relatively few design points so the distribution can be estimated with reasonable precision at a small number of places.
4. The choice of z_i that minimizes the confidence interval on β in non-linear models generally is not the one that minimizes confidence intervals on functions of α and β, and hence, the confidence interval for WTP.

As noted earlier, in the simplest case, estimates of mean (and median) WTP are a ratio of two parameters $(-\alpha/\beta)$, where α is the estimate of the constant from a logit or probit model and β is the estimate of the cost parameter. Two basic criteria are used in the stated preference literature for this case: 1. directly minimize the confidence interval around the mean WTP estimate and 2. maximize the determinant of the information matrix for the estimated parameters. Statistical designs that minimize confidence intervals around mean WTP are known as C-optimal designs. Alberini and Carson (31) and Alberini (32) show that C-optimality can be substantially more efficient (on the order of 50%) than maximizing the determinant of the information matrix (D-optimality) under conditions relevant to DCE studies.[25] Both the C- and D-optimality criteria lead to choosing only two design points if the underlying distribution can be fully characterized by two parameters and the design is not constrained to have more design points.[26] C- and D-optimal designs differ in where the points are placed, with D-optimal designs generally placing them further in the tails of the distribution.

[25] C-optimal designs are closely related to fiducial designs popular in biometrics.

[26] A design can be constrained to have more design points, but forcing a design to have four design points results in two design points being replicates, or being arbitrarily close to the two original points if they are forced to be distinct. If the distribution is assumed symmetric, equal numbers of observations generally are assigned to design points on either side of the median; asymmetric distributions can result in an asymmetric assignment of observations being optimal.

D-optimal designs are popular even in binary discrete choice cases with one z_p as a regressor, as it is natural to think in terms of maximum likelihood estimation; they also are easier to construct than C-optimal designs. D-optimal designs become more compelling in cases where goods are bundles of attributes and interest lies not in a single WTP estimate but in WTP estimates for a sizeable number of marginal tradeoffs. In this case, D-optimal designs effectively strike a balance in estimating all the marginal effects with reasonable precision for a given sample size. Much of the rest of this chapter is devoted to D-optimal designs when there are multiple attributes of interest.

Generally speaking, a D-optimal design is one that minimizes the determinant of the Fisher Information Matrix associated with a particular class of designs. The best D-optimal design is the one with the largest determinant. Street and Burgess (33) show that such designs exhibit level balance (each level of each attribute occurs equally often), and the differences in the attribute levels are orthogonal. It is difficult to generalize beyond this description because designs for non-linear models like choice models depend on the particular problem specification, namely the number of attributes, the number of levels associated with each attribute, the indirect utility specification associated with the problem, and the form of the underlying choice process model.

Both C- and D-optimality rely on certain knowledge of the model parameters, but this is never satisfied in practice because if the parameters were known there would be no need to do an experiment. Yet, one usually has at least some knowledge of the likely parameter values, so a good way to begin is to ask if theory can bound the parameter space, with inequality constraints being quite useful. Additionally, does existing literature on related goods help to bound the likely estimate of mean/median WTP? An obvious next step is to use data from pre-test and pilot studies to assist with this. Such a process is better thought of as 'sequential design', and Kanninen (34) discusses issues related to such a sequential design process. In general, the more uncertainty about the nature of the underlying WTP distribution, the more design points one should use, which can be shown using a formal Bayesian approach to design problems. Yet, one needs to recognize that there is a clear tradeoff between the precision at which the distribution is pinned down at individual design points and the number of design points.[27]

10.5 Generalizing attributes of binary discrete choice questions

Now, we consider what happens if an attribute is not continuous, but instead categorical, with greater than two levels. For example, an attribute of a GP practice might

[27] Alberini and Carson (31) suggest it is hard to justify more than eight design points, and show that four to six design points spanning the expected quartiles of the expected WTP give estimates that are reasonably efficient and robust to fairly large deviations in expected and observed WTP distributions, as long as the presumed distribution of WTP is of low dimensionality. McFadden (35) shows that a very different design is required if one wants to be able to consistently estimate mean WTP without making parametric assumptions about the nature of the distribution; this design involves spacing a large number of design points over the support of the WTP distribution.

be opening hours, such as 9–5, 9–7 and 9–9. If only this attribute is presented and the respondent's response options are to keep the status quo or choose the new opening hours, the theory and analysis are the same as in Section 10.4. Now, consider the case of adding a cost attribute with three discrete levels that cover the possible range of costs to opening hours. Now, we have a case where we must jointly vary two attributes each having three levels. The design involves both attributes; all combinations of them represent a 3 × 3 factorial design (= 9 combinations).

The implied DCE involves offering a respondent nine (or fewer) combinations of opening hours and costs. For each of the nine combinations that we will call 'scenarios', a respondent is asked whether they will stay with the status quo or switch to the new health service represented by a particular level of opening hours and a level of cost. Any design that uses less than all nine combinations will have some parameters that are not statistically identified without some identifying assumption/restriction for the underlying utility function.

If one believes that the true relationship between utility and cost is linear, the proper way to design this experiment is to only use two levels for the cost attribute, as discussed in Section 10.4. Thus, this DCE would have only six combinations. On the other hand, if one does not know the true relationship, and it is possible, perhaps likely, that it is non-linear, then one needs to assign *at least* three levels to the cost attribute. Typically, one would assign four levels to the cost attribute to be able to visualize relationships between utility and cost and rule out a quadratic polynomial if it is inappropriate. For example, if the true relationship is S-shaped, one needs at least four levels to visualize and test this.

The previous theoretical insights also apply to this case. That is, one may wish to value a change in opening hours from 9–5 to 9–9. This requires one to estimate the value of the utility difference between the two levels of opening hours, and if the relationship between utility and cost is linear, one would divide this utility difference by β, the estimate of the cost effect. If the status quo option varies across consumers, one must calculate the difference between the status quo and the proposed change in opening hours for each consumer and use the method of sample enumeration (36) to calculate the implied WTP. As before, if cost is treated as a random effect, one needs to calculate statistics for the WTP distribution, and one may need to simulate the WTP distribution in the case of complex models that allow random effects and covariances among effects and/or non-constant diagonal error variances and covariances.

Lancsar and Savage (42) discuss calculation of WTP for cases involving forced choice of one or more of the alternatives compared with the status quo, relying heavily on Hanemann's (21) discussion of issues involved in applying welfare ideas to discrete choice problems. For example, in the case of a simple binary choice model where the choice is between a constant status quo and a series of one-at-a-time designed choice options, one needs to examine the utility of each alternative and the probability of each alternative being chosen using 'expected utility'. The WTP expression for this case is

$$\frac{1}{\beta(\text{cos}\,t)}\left[\ln\sum_{i=1}^{n} e^{V_i^0} - \ln\sum_{i=1}^{n} e^{V_i^1}\right], \qquad (10.11)$$

where the cost effect is as previously defined, and $\ln \sum_{i=1}^{n} e^{V_i}$ is the so-called 'inclusive value', or expected maximum utility for the status quo (superscript 0) and the alternative of interest (superscript 1). So, (10.11) tells us that in cases where consumers can choose two or more alternatives one must evaluate the difference in expected utility between two options divided by the cost effect to calculate WTP. If the cost effect is non-linear, a more complicated expression is required but the concept is the same. If both cost and attribute are random effects, one must simulate the distribution.

Including more attributes is a direct extension of the above discussion. In general, for attributes $X_1(l_1)$, $X_2(l_2)$, …, $X_k(l_k)$, the total number of combinations is given by the full factorial expansion $X_1(l_1) \times X_2(l_2) \times \ldots \times X_k(l_k)$, where X_k is the kth attribute and l_k is the number of levels of that attribute. Thus, a DCE that involves asking a sample of respondents to compare a status quo option with some number of designed options one-at-a-time can be designed by (a) constructing the full factorial, and if sufficiently small, assigning all respondents to it, or if too large to do that, blocking the factorial into subsets (typically, randomly assigning sets without replacement) and assigning respondents randomly to each block (version); (b) using a fractional factorial design to sample from the full factorial and assigning all respondents to the scenarios given by the fraction, or blocking the fraction as described for the factorial, and randomly assigning respondents to one of the blocks (versions).

Random utility again underpins the specification of statistical models used to describe the choice process of respondents who participate in such DCEs. That is, as before we think of an indirect utility function with systematic and random components. Respondents seek to maximize their utility in their choices, but the analyst fails to include all factors known to the respondent and/or the respondent makes choices imperfectly, giving rise to the random utility case. Appropriate statistical models of the choice process for this case include (a) fixed effects for the attributes with additional terms that represent interactions of observable covariates with the intercept that reflects the propensity to choose the status quo versus the other options and/or interactions with the attributes or (b) random effects for the intercept and/or attributes to capture unobserved, latent differences in preferences (or, possibly a hybrid of a and b). Currently, random effects models are popular with academics and practitioners, but it remains unclear how to use such models to forecast choices and/or evaluate policies that will occur in the future and/or in other locations unless one assumes that random components are stable over time and/or space.

10.6 **Multinomial alternatives**

This case has two different versions: 1. there are multiple alternatives but all alternatives are generic and 2. at least one of the multiple alternatives differs in some significant way that requires the analyst to view this as 'non-generic' (or, 'alternative-specific'). For example, suppose a person's GP asks them to have a particular diagnostic test, and informs them that the test service is provided by (a) several named hospitals, (b) several named clinics, and (c) several named stand-alone testing services. Suppose further that the testing options can be described by (i) waiting time to be tested, (ii) locational convenience to the person, and (iii) cost.

If the objective is to understand and model the *type of service* that will be chosen by people facing this decision, or the *particular named option* within each type of service that will be chosen, then the problem is alternative-specific. If, on the other hand, the objective is to understand people's decisions/preferences for types or services and features of these services, where particular manifestations of the service options available for each type are examples, then the problem is generic. That is, alternative-specific problems arise when one wants to model the choices of particular named options that are members of a general class of options; generic options arise when one wants to model the choices of non-named options that lie within the general class. The former provides very specific information about the choices of particular options that would be of interest to – say – the owners of each type of option (e.g., owners of testing clinics); the latter provides very general information about the entire class of possible options. Tables 10.1a and b below illustrate two possible choice tasks for these cases using the testing example.

These cases are treated at length in (1), Louviere, Hensher, and Swait (hereafter 'LHS') as 'generic' and 'alternative-specific' DCEs. For generic DCEs, designs discussed and illustrated in LHS are obsolete because optimal design theory developed by Deborah Street, Leonie Burgess and colleagues (e.g., (33)) provides the theory and methods to construct optimal designs for this case. For 'alternative-specific' DCEs, the design theory originally proposed by Louviere and Woodworth (37) remains the primary way to construct such experiments. It is important to note that in the latter case, identification issues are well-understood and typically can be satisfied in virtually all

Table 10.1a An alternative-specific task

Features	Hospital	Clinic	Stand-alone
Wait time for testing	Same day	1 week	Next day
Locational convenience	15 min away	1 hour away	2 hour away
Cost	$75	$50	$100
I most likely will choose:	❑	❑	❑

Table 10.1b A generic task

Features	Option A	Option B
Type of test service	Hospital	Clinic
Wait time for testing	1 week	Next day
Locational convenience	2 hour away	15 min away
Cost	$100	$50
I most likely will choose:	❑	❑

applications but the efficiency of the designs relative to an optimal design is unknown.

Those who wish to construct optimally efficient designs for the generic case should consult Street and Burgess (33), which provides software to help analysts implement the design theory. In the case of alternative-specific designs, the theory detailed in LHS (1) or Louviere and Woodworth (37) provide the way to construct the designs. In the case of generic designs, attribute parameters specified in indirect utility functions are the same for all choice options, whether specified as fixed or random. In the case of alternative-specific designs, attribute parameters can be specified to be the same for some, but not all effects. That is, at least one attribute effect must differ for at least one alternative, regardless of whether the effects are specified as fixed or random.

Generic DCEs and associated models are consistent with the previous discussion of the theory that underlies calculation of WTP. If some model effects are alternative-specific, this implies that WTP will differ by (at least one) alternative. If cost effects are alternative-specific, this raises interesting issues over which cost effects to use in WTP calculations, as different cost effects imply different values of marginal values for income. Our position is that one generally should use cost effects associated with a given option to calculate the WTP associated with changes for that option, particularly when the cost for the base option is known. Comparisons between multiple programs are more complicated and, for this reason, researchers often try to use only one cost parameter unless there is clear evidence to the contrary.

A variation on the above theme is a DCE that presents respondents with multiple choice options, where one of these options is a constant option. To this point, the constant option has always implicitly been the status quo, but when multiple choice options are offered to respondents, a logical choice often is to choose none of the options, which is feasible since it involves zero cost. In the case of a constant status quo option, one can choose to incorporate the attribute levels of the status quo option in the estimation matrix or treat it as a fixed or random effect. In the case of the 'choose none' option, there are no associated attributes, and hence, one must be careful about how to specify this option. For example, as before, one can choose to allow it to be a fixed or random effect and/or one can allow the variance of the random component associated with this option to differ from the other options (as in nested or tree logit models).

10.7 Common designs

Two good sources of information about designs used in SP studies are LHS (1) and Street and Burgess (33). As these sources are available, our focus here will be on briefly describing the options and their advantages and disadvantages.

10.7.1 Ad hoc designs

From time-to-time one sees ad hoc designs used in SP studies. By 'ad hoc' we mean a design that is constructed without reliance on formal statistical design theory. Basically, one should ***NEVER*** do this, primarily because the properties of such designs are rarely known in advance, and it is likely that they are (a) statistically inefficient relative to an optimal design and/or (b) poorly conditioned, including the possibility of identification issues. Because applied economists rarely receive training in experimental design, and because econometrics historically has had to deal with 'messy' data, there is a tendency for applied economists to think that 'any design will do'. The sooner this notion is dispelled, the better.

10.7.2 **Full-factorial designs**

These designs may not be practical because the number of combinations of attribute levels can be very large. That said, let us distinguish two types of applications: 1. a design administered to everyone in a sample and 2. a design blocked into 'versions' with respondents randomly assigned to a particular version without replacement. Type 1 designs typically are used when one wants to be able to compare individuals and/or if one wants to estimate a model for each individual. For example, one type of comparison that often arises in practice is to group the individuals into segments based on their choices. The latter application is well beyond the scope of this chapter. Interested readers may wish to consult reference works on taxonomic methods such as cluster analysis or latent class methods. Type 2 designs typically are used when the design of interest has more than 16 or 32 attribute level combinations or choice sets and one does not want to compare individuals' choices directly.

Full factorials can be used as both type 1 and type 2 designs. In the case of type 1 designs, the class of factorials probably is restricted to those designs that have 32 choice sets or fewer, although it may be possible to use larger designs in certain cases where the incentives are sufficiently high, such as paying physicians enough to motivate them to 'do' perhaps 64 or more scenarios. In the case of multiple choice response tasks, however, only very small factorials are possible for type 1 applications. Researchers interested in such applications should consult Street and Burgess (33). In the case of type 2 applications, it is likely that full factorials can be practical for many cases because the factorial can be blocked into versions. For example, suppose that a researcher wished to design an experiment for ten attributes, each with two levels (2^{10}). The full factorial has 1,024 attribute level combinations, and if the researcher is confident that each respondent can and will 'do' 16 scenarios, the design can be blocked into 64 versions, with each respondent randomly assigned to one version. So, a full factorial of this size would be practical with samples of 400–800 people, which are not uncommon in SP applications. In our experience, many researchers rule out full factorials due to their size without realizing that they could have been used.

The major advantage of full factorials is that they allow one to estimate and test all possible main and interaction effects. In type 1 applications, there typically is a lot of statistical power to conduct these tests, and to the extent that one takes differences in individuals into account (e.g., preference heterogeneity), one can estimate and test these effects allowing for differences. The primary advantage of being able to estimate and test interaction effects is that one does not have to assume strictly additive indirect utility functions but instead can allow for more complex forms. The disadvantage, of course, is that there are typically many more effects to estimate, so analyses are more complicated. The advantage that is associated with type 1 applications does not necessarily apply to type 2 applications because (a) the power of the tests will be less due to smaller sample sizes associated with interaction effects, (b) it may not be possible to take individual differences into account as easily or as thoroughly as one can with type 1 applications because versions may be confounded with differences, and (c) it is unrealistic to rely on assumptions that all respondents have exactly the same indirect utility function.

In summary, full factorials probably can be used in many more applications than most SP researchers think, although their use in multiple choice response tasks is likely to be limited only to very small problems.

10.7.3 **Orthogonal main effects plans**

Orthogonal main effects plans (OMEPs) are a sample of attribute level combinations from the full factorial that have the property that all main effects are independent of one another. The advantage of OMEPs is that they typically are smaller than other designs. The disadvantage, however, is that one must assume that the indirect utility function is strictly additive for all respondents. Worse yet, if this assumption is false, one cannot test and reject it. It probably is fair to say that OMEPs are used much more often than they should, particularly in so far as they are the most widely used designs in SP research. If one has a type 1 design application, it may be that OMEPs are the only feasible option to allow rigorous comparisons of individuals. However, for type 2 design applications, it rarely would be the case that one would need to use an OMEP, and so researchers should consider other options discussed below.

10.7.4 **Designs that allow estimation of main and interaction effects**

Readers who want to construct and apply these designs should consult reference works in the design literature, although a good starting place is LHS (1). These types of designs are distinguished by what can be estimated and what must be assumed about omitted effects:

1. Main effects are orthogonal to one another, and are also orthogonal to unobserved but potentially significant two-way interactions. These designs protect estimates of main effects from two-way interaction effects that cannot be estimated. Their advantage is that they typically are relatively small(er), and hence, can be used in many applications. Two-way interactions are the most likely interactions to be significant and large, and so should be considered a key potential source of bias in main effects when they are omitted. Another disadvantage is that one must assume that all interactions of higher order than two-way are not significant, and that one cannot test this assumption to determine whether it is false. These designs typically need to be blocked into versions, but for smaller designs, it may be possible to use them in type 1 design applications.
2. One also can construct designs for problems that involve estimation of all main effects and a subset of the two-way interactions (known as 'selected two-way interactions'). It is hard to generalize about these designs because they typically are constructed on a case-by-case basis, but sources of these designs exist, as noted in Street and Burgess (33). The advantage of these designs is that they are smaller than

the design discussed below. The disadvantage is that designs for the exact subset that a particular researcher is interested in may not exist, and one must assume that all unobserved interactions are not significant, leaving one open to bias from unobserved two-way interactions that are significant.

3. Main effects and two-way interactions are orthogonal to one another, and both types of effects can be simultaneously estimated. These designs often are large, especially if the number of attributes and levels is greater than 8–10. However, for smaller problems, these designs have the advantage that they allow estimation of both main and two-way interaction effects, but at the cost of assuming that all other unobserved interaction effects are not significant. It rarely will be the case that one can use these designs in a type 1 design application, so the vast majority of applications of these designs will be for type 2 design problems.

4. One also can construct designs that allow one to orthogonalize the main effects and two-way interactions to unobserved and potentially significant three-way interactions. One also can construct designs that allow independent estimation of all main, two-way and three-way interactions. These types of designs are rarely used because they typically are fairly large, and require considerable design skill.

10.7.5 D-optimal designs

As previously noted, these designs optimize the determinant of the Fisher Information Matrix for the design. D-optimal designs for the case of all effects equal to zero have been developed by Street and Burgess (33), and readers should consult this reference for construction methods. What appears to be widely misunderstood about these designs is the fact that Monte Carlo simulations show that these designs are optimally efficient for choice probabilities that are not extreme and that they remain reasonably efficient for very large and very small choice probabilities. So, they are a good choice for almost all DCE problems. It also is worth noting that if one observes choice probabilities in the far tails in a DCE, this implies a very poor choice of the attribute levels or that the underlying process is almost deterministic. In addition to the Street and Burgess designs, designs can be constructed using SAS macros developed by Kufeld (38). Comparisons with Street and Burgess designs, however, suggests that the SAS designs sometimes do not have diagonal information matrices and can require substantial computation time to construct a highly efficient design.

10.7.7 Random designs

A number of SP researchers use what we call 'random designs'. These designs are constructed in various ways, but typically one or more sets of starting designs are constructed or one randomly samples from the complete set of all possible choice sets. If one uses a set of starting designs, say *m* of them, one typically randomly selects an attribute level combination from each of the *m* simultaneously to create an *m*-tuple that represents an *m*-element choice set. This design procedure was discussed by Louviere and Woodworth (37), but modern advances in optimal design of DCEs has made them obsolete. Similarly, some commercial DCE software creates choice sets by drawing them randomly from the entire set of possible choice sets, which typically is very large. Neither way of designing DCEs is a good idea since one cannot determine *a priori* which effects can be estimated with any precision, nor can one identify *a priori* what will be identified.[28] This approach also has the disadvantage that differences in individuals may be confounded with differences in the choice sets faced. Our advice is not to use this approach since better alternatives are available.

10.8 **Using prior information**

Naturally, it is always better to use whatever prior information one has available to construct designs, as noted earlier. So, if one has theoretical or empirical reasons to impose sign restrictions on the attribute effects, this will (a) restrict the classes of designs to consider and (b) will restrict the indirect utility functions to be estimated. For example, suppose there are three two-level attributes, and each has a known sign for the main effects. If the indirect utility function is strictly additive, only four scenarios are required to estimate the model in a binary discrete choice task. There are 16 possible binary response patterns that could be observed because each of the four scenarios can receive either of the two binary responses (2^4). Of these 16 patterns, only seven are consistent with additivity and sign restrictions, assuming that basing responses on only a single attribute or a pair of attributes is acceptable.

If the utility function is not additive, one must use the full factorial ($2 \times 2 \times 2 = 8$ attribute level combinations), which greatly increases the allowable response patterns to nearly 128, again assuming that one can base one's choices on one or a pair of attributes as well as all three attributes. It should be obvious that as the number of attributes and/or the number of response categories increase, the number of possible response patterns that can be consistent with a particular set of sign restrictions grows exponentially. Thus, as the number of attributes and/or the number of levels increases, sign restrictions may not help bound the problem in any practical sense. As such, sign restrictions are most useful for smaller problems.[29]

Finally, one can use a sequential design approach. In this approach, one uses experiments on small(er) samples to explore as much of the design space as possible. This approach can be viewed as a type of model selection problem where the objective is to identify as many possible significant and meaningful effects as possible *a priori*, while at the same time eliminating as many non-significant and non-meaningful effects as possible. In this way, one can bound the problem, which may allow one to use a smaller, special purpose design to identify and estimate the effects that one has *a priori* reason to believe will be significant and meaningful.

[28] Use of random design is often believed to identify all of the parameters of a model. That is true, however, only asymptotically as the design drawn approaches the full factorial. In the typical application, many parameters will not be statistically identified and other very poorly identified if subsamples receiving particular attribute combinations is small.

[29] One also can impose informative priors on the parameters of the utility function. This takes one in the direction of Bayesian designs if uncertainty around the priors is formally quantified. If one is prepared to assume that the parameters are known with certainty, it is possible to determine the design that maximizes D-optimality.

10.9 Desirability and implications of common design criteria

In this section, we review several design criteria that are frequently discussed in the various DCE literatures.[30]

10.9.1 Orthogonality

Orthogonality of the effects to be estimated means that the information matrix is block diagonal, and hence, all effects of interest can be estimated independently of one another. This is a desirable but not essential criterion. What is essential is that the degree of shared covariance between effects to be estimated is low.

10.9.2 Level-balance

This means that each level of an attribute occurs equally often, and more generally, this should hold for all attributes. This criterion is associated with the precision of the estimate of the attribute levels, such that if level balance holds, the model parameters associated with each level will be estimated with equal precision. Again, this criterion is desirable, but not essential. However, unless one has good reasons for not satisfying level balance, such as one of three levels being far more important to the work than the other two, it is desirable to satisfy this criterion. A similar criterion is a balanced level co-occurrence. That is, if a design is orthogonal, it will be the case that the levels of each pair of attributes will co-occur equally often. This criterion insures minimal shared covariances.

10.9.3 Attribute overlap

This refers to correlations among two or more attributes, such that it may not be possible to vary them independently. This leads to what are called 'nested' factors/attributes because the way to deal with these problems is to combine the attribute levels into a single attribute that can be varied independently. For example, if a particular health service attribute is the amount of use of the system, and a second is the cost of using the system, it is likely to be the case that higher levels of use will covary with cost, so one would want to combine these two attributes into one.

10.9.4 Elimination of dominated/infeasible alternatives

If all the attributes are numerical and their signs are known *a priori*, any of the standard design constructions will lead to dominated and/or infeasible options. However, while this happens in practice, our experience suggests that there is far too much concern about this criterion than should be the case.[31] The first thing a researcher should

[30] Viney, Savage, and Louviere (39) look at these concepts in the context of a specific empirical example.

[31] In empirical applications, the more serious problem is likely to be that particular attributes are known to have sufficiently high correlation so that the absence of this correlation in attribute bundles is noticeable.

do is to determine whether it is reasonable to expect the respondent sample to actually know which attribute level combinations are infeasible. Typically, the sample does not know this; but the experts, such as doctors or medical researchers do. If the sample cannot tell if a scenario or an option is infeasible, a researcher may want to proceed to use standard design construction methods as these will provide significantly better statistical properties than alternative methods.

If many respondents in fact know that something is infeasible and/or there can be dominant options in choice sets, then one typically must modify the standard construction methods to deal with this. For example, one way to deal with dominance is to randomly replace one or more levels with levels that are non-dominant. Ideally, one should test various random replacements, using those that minimally modify the statistical properties of the design. Infeasible options are a different problem as they have to be eliminated from the design. One way to do this is to construct the full factorial of possible options that can be created from the attributes and levels of each choice option (if the options are not generic), then eliminate all the infeasible combinations and check the statistical properties of the remaining combinations. If the shared covariances are not large and the inverse of the information matrix is well-conditioned, then use that design. If the statistical properties are poor, then one could try to select combinations from the feasible pool with the objective being to select a sample that has the best properties.

10.9.5 Utility balance

It is not clear why this criterion is considered important, although it has achieved considerable prominence in the marketing literature. Put simply, this criterion means that one should try to construct choice sets in such a way that the options in each set are as close in utility as possible. While this may seem like an intuitive criterion, if one could achieve this objective, there would be ***NO*** useful statistical information provided by the choices. That is, satisfying this criterion is equivalent to making all the choice options equally probable because if the option utilities are perfectly balanced, the respondents should be totally indifferent to all of them, and so should choose randomly. Thus, this criterion should not be used in the design of DCEs.

10.10 Concluding remarks

Experimental design is a key component of a successful choice experiment to help evaluate health policy alternatives. It is all too easy to construct and implement designs that do not statistically identify the parameters of interest or that greatly diminish the precision of the estimates relative to what could have been achieved with an efficient design. The underlying statistical theory for generic choice experiments is now well-understood (33), and software for producing reasonably high quality designs is now available (e.g., Street and Burgess design software that comes with their book); so, there is little justification for choosing and using the poor quality designs that appear all too often in the current literature.

Applied researchers need to think seriously about the attributes of the programs they wish to compare and the class of underlying utility functions they wish to estimate.

Invest time upfront in extensive qualitative work of the type illustrated by the ICEPOP Program (31;41), where extensive and iterative qualitative work was used to understand not only attributes, but also key words and phrases. Also plan time for elaborate and extensive pre-testing to identify tasks and associated survey instruments that 'work'. By 'work' we mean that (a) will be understood by all respondents, (b) will be meaningful to them, and (c) will simulate the actual choices one wants to observe as closely as possible. Often it makes sense to ask more choice questions rather than more complicated choice questions, or to ask more questions about the options in each choice set instead of more choice sets. Make reasonable restrictions on the nature of the utility function to reduce the size of the model space. Fix the attribute levels that are relatively unimportant to the policy issues being evaluated to further reduce the size of the model space and the task. More generally, one should avoid complex models and designs unless the project budget allows for extensive pretesting and large sample sizes.

Spend time observing the actual choices made by the population(s) of interest. Interview these populations and ask them how they make the choices, what are the pros and cons of each choice, and whether they feel like they have sufficient and/or 'the right' information to make the right decision(s); if they do not have sufficient or 'right' information, what would assist them? Relying on experts to tell you how consumers make decisions rarely is a good idea because few of them actually know this. If, in fact, the expert is the one who makes the decision for the consumer, then model the expert; yet, even here, one may want to understand and model how the expert's decision(s) impact(s) the consumer.

Finally, there is a great deal of misinformation and misunderstanding about the design of DCEs in the applied health economics literature. The literature on the optimal design of DCEs is highly technical, and it is easy to make mistakes as noted by (33;43). There are no quick fixes and no easy routes; the literature on the design of experiments for linear models has evolved over more than 80 years, with some problems yet to be resolved. So, beware those who claim to have answers for all DCE design problems. Currently, we barely understand the generic design case for conditional logit models, and while designs for the alternative-specific case have been around since Louviere and Woodworth (37), few formal proofs of the properties of these designs exist even for conditional logit models. Furthermore, there are virtually no results available to guide those who want to estimate more general choice models than conditional logit, although mixed logit models at least should be identified with current generic and alternative-specific designs (each person is represented by a conditional logit model; only the parameters of that model differ across people).

It also is worth noting that this chapter has had little to say about types of tasks, task context, task complexity, methods of survey administration, survey length, incentive compatibility, sampling strategies, and a host of other issues relevant to whether any given stated preference DCE survey is reliable and valid. Similarly, we have said nothing about validating SP model predictions, pooling data from various sources, taking account of observable and unobservable heterogeneity and many other issues that are germane to particular applications. The choice modelling and SP literatures are now extensive on each of these topics, and interested readers should consult the workshop reports from the triennial Invitational Choice Symposia published in special issues of the journal Marketing Letters (e.g., Marketing Letters 1991, 1993, 1996, 1999, 2002, 2005, 2008), as well as standard reference sources like (1;44) for guidance.

References

1. Louviere, J.J., Hensher, D.A., and Swait, J.D. 2000. *Stated choice methods: analysis and application.* New York: Cambridge University Press.
2. Bradburn, N.M., Sudman, S., and Wansink, B. 2004. *Asking questions: The definitive guide to questionnaire design—for market research, political polls, and social and health questionnaires.* San Francisco, CA: Jossey-Bass.
3. Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., and Singer, E. 2004. *Methods for testing and evaluating survey questionnaires.* New York: Wiley.
4. Tourangeau, R., Rips, L.J., and Rasinski, K. 2000. *The psychology of survey response.* New York: Cambridge University Press.
5. Mitchell, R.C. 2002. On designing constructed markets in valuation surveys. *Environmental and Resource Economics* **22**: 297–321.
6. Kaplowitz, M.D., Lupi, F., and Hoehn, J.P. 2004. Multiple methods for developing and evaluating a stated choice questionnaire to value wetlands. In: Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., and Singer, E., editors. *Methods for testing and evaluating survey questionnaires.* New York: Wiley.
7. Swait, J. and Adamowicz, W. 2001. The influence of task complexity on consumer choice: A latent class model of decision strategy switching. *Journal of Consumer Research* **28**: 135–148.
8. Viney, R., Lanscar, E., and Louviere, J. 2002. Discrete choice experiments to measure consumer preferences for health and healthcare. *Expert Review of Pharmaco-economics and Outcomes Research* **2**: 319–326.
9. Ryan M. and Gerard, K. 2003. Using discrete choice experiments to value health care: current practice and future prospects. *Applied Health Economic Policy Analysis* **2**: 55–64.
10. Bryan, S. and Dolan, P. 2004. Discrete choice experiments in health economics: for better or worse. *European Journal of Health Economics* **5**: 199–2002.
11. Lancsar, E. and Donaldson, C. 2005. Discrete choice experiments in health economics: distinguishing between the method and its applications. *European Journal of Health Economics* **6**: 314–316.
12. Viscusi, W.K. and Gayer, T. 2005. Quantifying and valuing environmental health risks. In: Karl-Göran, M. and Jeffrey R.V., editors. *Handbook of Environmental Economics*, vol. 2, Amsterdam: North-Holland.
13. Mitchell, R.C. and Richard, T.C. 1989. *Using surveys to value public goods: The contingent valuation method.* Baltimore, MD: Johns Hopkins University.
14. Arrow, K., Solow, R., Portney, P.R., Leamer, E.E., Radner, R., and Schuman, H. 1993. Report of the NOAA Panel on Contingent Valuation. *Federal Register* **58**: 4601–4614.
15. Carson, R.T. and Groves, T. 2007. Incentive and information properties of preference questions. *Environmental and Resource Economics* **37**: 181–210.
16. Just, R.E., Darrell, L.H., and Andrew, S. 2005. *The welfare economics of public policy: A practical approach to project and policy evaluation.* Northhampton, MA: Edward Elgar.

17. Bockstael, N.E. and Freeman, A.M. 2005. Welfare theory and valuation. In: Karl-Göran, M. and Jeffrey, R.V., editors. *Handbook of environmental economics*. vol. 2, Amsterdam: North-Holland.
18. Carson, R.T. and Hanemann, W.M. 2005. Contingent valuation. In: Karl-Göran, M. and Jeffrey, R.V., editors. *Handbook of environmental economics,* vol. 2, Amsterdam: North-Holland.
19. McConnell, K.E. 1990. Models for referendum data: the structure of discrete choice models for contingent valuation. *Journal of Environmental Economics and Management* **18**: 19–34.
20. Cameron, T.A. 1988. A new paradigm for valuing non-market goods using referendum data: maximum likelihood estimation by censored logistic regression. *Journal of Environmental Economics and Management* **15**: 355–379.
21. Hanemann, W.M. 1984. Welfare evaluations in contingent valuation experiments with discrete responses. *American Journal of Agricultural Economics* **66**: 332–341.
22. McFadden, D.L. and Gregory, K.L. 1993. Issues in the contingent valuation of environmental goods: methodologies for data collection and analysis. In: Jerry, A.H., editor. *Contingent valuation: A critical assessment*, pp. 165–216. Amsterdam: North-Holland.
23. Lancaster, K. 1966. A New Approach to Consumer Theory. *Journal of Political Economy* **84**: 132–157.
24. Carson, R.T. and Yongil, J. 2000. On overcoming informational deficiencies in estimating willingness to pay distributions, paper presented at the American Agricultural Economics Association Meeting, Tampa, FL.Chilton, S.M., and W.G. Hutchinson (1999), Do focus groups contribute anything to the contingent valuation process? *Journal of Economic Psychology* **20**: 465–483.
25. McFadden, D.L. 1974. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. editor. *Frontiers in econometrics*, pp. 105–142. New York: Academic.
26. Manski, C. 1977. The structure of random utility models. *Theory and Decision* **8**: 229–254.
27. Kriström, B. 1997. Spike models in contingent valuation. *American Journal of Agricultural Economics* **79**: 1013–1023.
28. Box, G.E.P., Hunter, W.G., and Hunter, J.S. 1978. *Statistics for experimenters: An introduction to design, data analysis, and model building.* New York: Wiley.
29. Atkinson, AC. and Donev, A.N. 1992. *Optimum experimental designs.* New York: Oxford University Press.
30. Wu, C.F.J. and Hamada, M. 2000. *Experiments: planning, analysis, and parameter design optimization.* New York: Wiley.
31. Alberini, A. and Richard, T.C. 1990. Choice of thresholds for efficient binary discrete choice estimation. Discussion Paper 90-34. San Diego: Department of Economics. University of California, San Diego.
32. Alberini, A. 1995. Optimal designs for discrete choice contingent valuation surveys: Single-bound, double-bound, and bivariate models. *Journal of Environmental Economics and Management* **28**: 287–306.
33. Street, D.J. and Burgess, L. 2007. *The construction of optimal stated choice experiments: theory and methods*. New York: Wiley.
34. Kanninen, B.J. 1993. Design of sequential experiments for contingent valuation studies. *Journal of Environmental Economics and Management* **25**: S1–S11.
35. McFadden, D.L. 1999. Computing willingness-to-pay in random utility models. In: Moore, J., Riezman, R., and Melvin, J., editors. *Trade, theory, and econometrics: essays in honor of John S. Chipman*. London: Routledge.

36. Ben-Akiva, M. and Lerman, S. 1985. *Discrete choice analysis: theory and application to travel demand.* Cambridge: MIT.

37. Louviere, J.J. and Woodward, G. 1983. Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data. *Journal of Marketing Research* **20**: 350–367.

38. Kufeld, W. 2005. Experimental design and choice modeling macros. Technical Report TS722I, Cary, NC: SAS Institute.

39. Viney, R.,Savage, E., and Louviere, J. 2005. Empirical investigation of experimental design properties of discrete choice experiments in health care. *Health Economics* **14**: 349–362.

40. Coast, J., Flynn, T.N., Natarajan, L., Sproston, K., Lewis, J., Louviere, J.J., and Peters, T.J. 2008. Valuing the ICECAP capability index for older people. *Social Science and Medicine* **67**(**5**): 874–882.

41. Coast, J., Flynn, T.N, Sutton, E., Al-Janabi , H., Vosper, J., Lavender, S., Louviere, J.J., and Peters, T.J. 2008. Investigating choice experiments for preferences of older people (ICEPOP): evaluative spaces in health economics. *Journal of Health Services Research & Policy* **13**(3): 31–37.

42. Lancsar , E. and Savage, E. 2004. Deriving welfare measures from discrete choice experiments: inconsistency between current methods and random utility and welfare theory. *Health Economic Letters* **13**: 901–907.

43. Street, D.J., Burgess, L., and Louviere, J.J. 2005. Quick and easy choice sets: constructing optimal and nearly optimal choice experiments. *International Journal of Research in Marketing* **22**: 459–470.

44. Train, K. 2003. *Discrete choice methods with simulation*. New York: Cambridge Economic Press.