A review of the discrete choice experiment - with emphasis on its application in health care

Trine Kjær
Health Economics
University Of Southern Denmark

Health Economics Papers
2005:1

1 Introduction 1

1.1 Objective and structure of the paper 3

2 CBA and preference-based outcome measurement 5

2.1 The theoretical foundation of CBA 6

2.2 Economic value 7

2.3 Stated versus revealed preference methods 10

2.3.1 The hedonic pricing method 13

2.3.2 The travel cost method 14

2.3.3 The use of revealed preference methods in health economics 15

2.4 The contingent valuation method 15

2.4.1 CVM in health economics 18

3 Fundamentals of the discrete choice experiment 19

3.1 The discrete choice experiment 22

3.1.1 Inclusion of a cost attribute 24

4 Foundation of the discrete choice experiment 27

4.1 Random decision rule 28

4.2 Random Utility Theory 31

4.3 The utility function 34

4.3.1 Including interaction terms 36

4.4 Discrete choice modelling 37

4.4.1 Binary discrete choice models 41

4.4.2 Extension of binary models: Random effects 43

4.4.3 Multinomial discrete choice models 44

4.4.4 Alternative specific constants 56

4.4.5 Estimation procedure – the likelihood function 57

4.4.6 Combining of data - and scaling 59

4.4.7 Measuring welfare 66

4.4.8 Derivation of standard errors of MRS and WTP 73

5 Design of the DCE 75

5.1 Stage 1: Identification of attributes 76

5.2 Stage 2: Identification of levels *79*
5.2.1 Level range *79*
5.2.2 Attribute-effect *80*
5.3 Inclusion of a cost attribute *81*
5.4 Stage 3: Experimental design *84*
5.4.1 Factorial design *84*
5.4.2 Ensuring high design efficiency *86*
5.5 Complexity versus completeness - The extent to which design influences cognitive demand *90*
5.6 Stage 4: Questionnaire construction and data collection *93*
5.6.1 Inclusion of an additional alternative: the opt-out/status-quo *93*
5.6.2 The introductory text *95*
5.6.3 Presentation of the choice sets *96*
5.6.4 Inclusion of a validity test in the choice task *97*
5.6.5 'Uncertainty question' and 'cheap talk' *98*
5.6.6 Follow-up questions *99*
5.6.7 Socio-demographic questions *100*
5.6.8 Focus group *100*
5.6.9 Data collection method *101*
5.6.10 The sample *103*
5.7 Stage 5: Data analysis *104*
5.7.1 Data input *104*
5.7.2 Data analysis *105*
6 Biases and validation **106**
6.1 Biases *107*
6.2 Validity and reliability *110*
6.2.1 Content validity *111*
6.2.2 Criterion validity *111*
6.2.3 Construct validity *112*
6.3 CVM versus DCE *117*
7 Conclusion **120**

References **122**

Appendix I. Illustrations of welfare measures **136**

Appendix II. Illustrations of coding and data input **137**

Appendix III. An example of a regression analysis **139**

1 Introduction

Increased spending on health care, together with a scarcity of resources, has led to greater focus on priority-setting within health care and hence more frequent application of economic evaluation. To date the most widespread evaluation methods used in health care are cost-effectiveness analysis (CEA) and cost-utility analysis (CUA). Common to these evaluation methods is that they examine the effect of an intervention and the decision-making rule is to optimize effect per cost. In CEA the effect is a one-dimensional measure such as blood pressure, life year saved, etc., whereas in CUA it is a multidimensional measure in which the quality of the life year saved (i.e. 'quality-adjusted life year' or QALY) is taken into consideration. An important feature of these health outcomes is that they only allow for health-related preference-based outcome measures, meaning that only health-related measures of benefits are considered. There has been an assumption in the health economic literature that such health outcomes are all that needs to be addressed when estimating the benefits from health care interventions (Ryan 1999b). The decision rule underlying CEA and CUA thus built on a narrow concept of utility compared to the definition of utility known from neoclassic economic theory; neoclassical economics being the theoretical foundation of the third type of evaluation method known as the cost-benefit analysis (CBA). CBA aims to maximize aggregated welfare, where 'welfare' constitutes all the elements that provide individuals with utility. Compared to CEA/CUA, therefore, CBA takes into account whatever preferences individuals have. The CBA approach makes it possible to consider individuals' utility arising from non-health outcomes, psychological feelings such as regret and disappointment, and process utility (Drummond et al. 1997;Ryan 1996). Moreover, Ryan (2004c) notes that applying the CBA approach may well lead to conclusions that conflict with the recommendations of CUA. Whilst CBA and partly CBA (monetary elicitation of individuals' preferences) have grown in popularity amongst some groups of academic health economists, it is less well-known among policy makers and the pharmaceutical industry. Policy recommendations from the National Institute of Clinical Excellence (NICE) in England, for example, favour the use of CUA and CEA in health care prioritizing (Hanley et al. 2003;Hutton & Maynard 2000). In environmental policy-making, however, CBA has gained acceptance as the leading method of environmental evaluation and it is generally acknowledged that valuation of environmental goods should be based upon consumer preference (Hidona 2002). This may be due to the longer history of CBA application in environmental science, and also that useful alternative environmental economic evaluation techniques are limited.

It is argued that the CBA is the only economic evaluation method that is grounded in neoclassical welfare economics. It involves the valuation of benefits in the same unit as costs (i.e. monetary valuation), an approach that is required for advising decision-makers on improvements in allocative efficiency (Olsen & Smith 2001). To measure the benefit from a programme, the researcher measures how much the individual is willing to give up in order to get the benefit – this is known as willingness-to-pay (WTP). While some researchers advocate the use of WTP methods, others are more sceptical and prefer the application and development of other evaluation methods such as CUA (Cookson 2003). The controversy stems, among other things, from the fact that the two types of evaluation methods (CBA versus CEA/CUA) represent different sets of value judgements. As just mentioned, the principle in CBA is the maximization of the aggregated sum of individuals' utilities (referred to as 'welfarism'), whereas the principle in CEA/CUA is the maximization of society's health status (referred to as 'extra-welfarism'). Which of the two approaches is preferred depends on one's own beliefs. Olsen & Smith (2001) argue that the 'rightness' of an evaluation approach is to be judged on the basis that its value judgements are compatible to those society holds for the health care institution for which the particular economic evaluation is being undertaken. These value judgements differ between countries and between health care institutions within a country and hence there is no universally "correct' economic evaluation approach. As CBA is the only method that is founded in neoclassical economic theory, however, one could argue that this gives it an advantage. During the last decade, attempts have been made to link CUA with CBA such as finding a monetary measure for a QALY, a cost-per-QALY (e.g. Gyrd-Hansen 2003). Such a link seems appealing since the results of CUAs could then be interpreted within a standard welfare economic framework. However, for a unique measure of WTP-per-QALY to hold, a linear relationship between WTP and QALY has to exist; this requires some strong (and potentially unrealistic) assumptions regarding the utility function (Dolan & Edlin 2002).

The popularity of economic evaluation is driven by the fact that use of ordinary market analysis is often impossible in health care as the market either fails to work perfectly[1] or is absent. An example of a non-marketed good is hospital treatment, which is paid through taxes and hence free of charge for the individual. An example of a market failure (due to government intervention) is the market for pharmaceuticals in which the consumer only partly pays for the medicine (a co-

[1] Reasons for the market to fail in health care include well-known arguments such as imperfect competition, public goods theory, externalities, market imbalances and asymmetric information. Market failure leads to divergence between market prices and marginal social costs (Boardman et al. 2001;Johansson.P. 1991).

payment system). In this case the price at the market is not the equilibrium market price and hence the demand curve is unusable as an instrument in priority settings. In a perfectly competitive market, under certain conditions, the equilibrium price indicates both the marginal social costs and the marginal social benefits of the production of one more unit of that good. This is because opportunity costs of production are given by the supply curve whilst the demand curve is a schedule of marginal willingness-to-pay (Hanley & Spach 1993).

The measurement of benefit in a CBA for goods in which the market is absent or fails to work is not straightforward and involves a large number of considerations. One is the choice of method to elicit preferences. A variety of methods exist that have this potential - the discrete choice experiment being one of them.

1.1 Objective and structure of the paper

The objective of this paper is to provide the reader with a review of the discrete choice experiment (DCE) – with emphasis on its application in health care.

To understand and apply the method it is crucial to place it into the context of preference-based economic evaluation in general – this is done in chapter 2. Of special importance is the relation between the discrete choice experiment and the contingent valuation method, as these two methods share many features but also have essential differences. Chapter 3 introduces the fundamentals of the discrete choice experiment, and summarizes its history – starting at the beginning of the 1970s when the method was developed in the marketing literature as a market research tool for evaluating consumer behaviour and predicting sales of new products. When applying the discrete choice experiment in the field of health economics (or any other field of economics), it is essential to understand the theoretical foundation of the method. The discrete choice experiment is founded in random utility theory (RUT) and is consistent with Lancaster's theory of characteristics and neoclassic economics. RUT plays a key role in the understanding and interpretation of the behavioural processes examined in the DCE. Chapter 4 describes the theoretical foundation of the DCE and places it in the context of probabilistic theory and modelling of discrete choices. The application of a DCE requires many issues to be taken into consideration, such as the perspective of the DCE, decisions about attributes and levels, experimental design, data collection and so forth. These design issues are discussed in chapter 5. The DCE belongs to the class of stated preference methods which implies that actual individual behaviour is not observed,

but instead individuals are asked to consider a hypothetical scenario. Validation of the approach is therefore crucial – the researcher has to be sure that the method measures what it is intended to measure. Any divergences between actual, hypothetical and theoretical behaviour can be ascribed to biases. Validation and biases are discussed in chapter 6 along with the advantages and disadvantages of using the DCE compared to contingent valuation. Chapter 7 outlines present and future research areas and concludes the paper.

2 CBA and preference-based outcome measurement

The cost-benefit analysis (CBA) describes an analysis, which seeks to quantify in money terms the costs and benefits of a policy intervention or project. Basically the decision rule rests on the sign (and magnitude) of the estimated net economic value; i.e. benefits (economic value) minus opportunity costs. If net economic value exceed zero, the intervention/project should be implemented. The CBA approach was first developed in the 1950s to assess the net economic value of public projects that used productive factor inputs (e.g. land, labour, capital, materials, etc.) to produce tangible outputs. Many of the outputs had market counterparts, so estimation of monetary values was relatively straightforward. For example, the savings in the monetary costs of repairing flood damages measured the benefits of controlling floods. In contrast, the effects of many public actions today are much more subtle and wide-ranging. What were once considered unquantifiable and perhaps relatively unimportant intangibles are now recognised as significant sources of value and are considered to be suitable for economic measurement; consequences that were once unrecognised or were thought to lie outside the realm of economic analysis are often central issues in current policy analysis (Freeman 1999). According to Hidona (2002), the definition of valuation is the measurement of the value of goods or benefits of the projects and policies implemented. In terms of benefits, the value is the increment of utility minus the increment of disutility due to the implementation of the public project or policy that is measurable by the welfare measures.

If money is used as the standard to measure welfare, then the measure of benefit is the *willingness-to-pay* (WTP) to secure that benefit, or the *willingness-to-accept* (WTA) compensation. History shows that there is often a disparity between WTA and WTP measures (WTA being larger than WTP). Explanations for such disparity are many and include arguments such as substitution effect (Hanemann 1991)[2]; loss aversion[3] (Tversky & Kahneman 1991); uncertainty, irreversibility and learning effect (Zhao & Kling 2001) and bias. There is general agreement among economists, however, that the measuring of benefits should be performed in a manner in which individuals are asked to state their WTP as this measurement seems to be the most reliable and if anything is an underestimate of the true value (Arrow et al. 1993).

[2] The only argument aligned with neoclassical economic theory
[3] Also known as endowment effect and reference-dependent preferences

2.1 The theoretical foundation of CBA

The measurement of benefits (and costs) underlies the concept of economic efficiency. In economic theory, social welfare is measured through the weighted sum of individuals' utilities. In order to measure a change in welfare due to a policy change, it is therefore necessary to measure the changes in the consumers' utility. In order to be able to say anything about social welfare, value judgements have to be made. Whenever one situation is said to be better than another situation, the evaluation is based on a certain set of value judgements. The basic value judgement used in welfare economics is the *Pareto principle* (Boardway & Bruce 1984). This principle says that an allocation of resources is Pareto efficient if it is impossible to make somebody better off without making anyone else worse off. Thus Pareto optimality refers to a Pareto optimal state in which no re-allocation of resources is possible that makes somebody better off without making anyone else worse off. The Pareto principle has clear limitations as the principle is based only on the ordinal properties of the utility function (Johansson.P.O. 1991). Ordinal measures of utility indicate the direction, but not the magnitude, of change. Nor does such an evaluation facilitate a comparison of the changes in welfare across individuals, which is particularly important when a policy change makes some individuals better off and others worse off which is the case for real-world projects as these always produce both gainers and losers. In order to undertake such welfare analyses, it is necessary to have a monetary (cardinal) measure of the change in welfare, which is based on the concept of a *potential Pareto improvement* – also known as the Kaldor-Hicks criterion and the compensating principle – developed by Hicks and Kaldor in 1939. In such cases the application of the principle can be used as a decision criterion. According to the principle proposed by Kaldor, state *a* is preferable to state *b* if, in state *a,* it is hypothetically possible to undertake costless lump sum redistribution and achieve an allocation that is superior to the other state according to the Pareto criterion. Hicks offered a slightly different criterion, in which state *a* is preferable to another state if, in state *b*, it is not possible, hypothetically, to carry out lump-sum redistribution so that everyone could be made as well-off as in state *a* (Boardway & Bruce 1984). This means that state *a* is preferred to *b* if the welfare measure is above zero.

The controversial aspect of the Kaldor-Hicks principle is the hypothetical and lump sum nature of the redistribution. If the redistribution was actually carried out, the entire situation would be a direct use of the Pareto criterion itself. The purpose of considering hypothetical redistribution is to try and separate the efficiency and equity aspects of the policy change under consideration. It can be argued that whether or not the redistribution is actually carried out is an important but separate discussion

(Johansson.P.O. 1991). The fact that it is possible to create potential Pareto-improving redistribution is sufficient to rank one state above another on efficiency grounds. Moreover, it is often impossible to identify the losers and gainers in practice and thus to implement the compensation. The thought behind the principle is that gainers and losers differ from state to state such that, from a broad perspective, each individual experiences both gains and losses. Aggregating welfare across individuals remains a contentious issue. Applying the potential Pareto improvement criterion, it is a sufficient condition for a policy to improve welfare if the aggregated welfare measure is greater than zero so that the gainers could *potentially* compensate the losers. The elicitation of WTP is biased in favour of the existing distribution of income, since it assumes the initial distribution of income is appropriate. This has clear distributional consequences and hence is an equity concern (Johansson 1993). A possibility is to perform equity weighting by social classes such as income. However, such weighting is not usually incorporated in modern CBAs. Rather, equity is better achieved by looking at the overall set of policies and projects and making the appropriate adjustments via the tax and social security systems (Bateman et al. 2002).

2.2 Economic value

The economic concept of value has been generally defined as any net change in the welfare of society. From an economic perspective, values can be associated equally with the consumption of goods and services purchased in markets and with the services from goods and services for which no payments are made. In this sense, anything from which an individual gains satisfaction is deemed to be of value, so long as the individual is willing to give up scarce resources for it. The total economic value comprises explicit use benefits as well as implicit non-use benefits. In the There are therefore different types of economic value, in which the sum of all the values (i.e. WTPs) defines the total economic value of any change in wellbeing due to an intervention; see Figure 2.1. In this sense total economic value does not include the opportunity costs [4]associated with an intervention Besides the standard value associated with consumption of a good – known as 'use value' – it has been argued that individuals place monetary value on goods that is independent of their present consumption. This implies that the total economic value goes beyond what is possible to estimate using market analysis. The first systematic attempts to account for values other than 'use-value' were by Weisbrod (1964) and Krutilla (1967). Weisbrod focused on uncertainty

[4] Opportunity cost (also termed 'economic cost'): The value of the resource in its most valuable alternative use.

and what became known as ‘option value’, whereas Krutilla focused on ‘non-use value’ (also named ‘passive use value’). ‘Option value’ refers to the value that people place on having the opportunity to consume the good in the future. Due to the uncertainty of consumption linked to option value, there is disagreement among economists as to whether it should be classified as a use or a non-use value, or placed horizontal to these two values. The uncertainty paradigm related to option value has created problems due to the fact that future decisions are random as viewed from today. Since the publication of Weisbrod’s article, there has been much debate, especially among environmental economists, regarding the appropriate measure in such situations. It has been argued that when demand, for instance for a national park, is uncertain then the expected consumer surplus will differ from the constant maximum payment (option price) that the consumer is willing to pay across states. The difference arises because the option price, which is non-stochastic or state independent, measures both the value of retaining an option to consume the good and the expected value of actually consuming the good, i.e. expected consumer surplus. It is expected that this difference, named option value, would be positive for risk-averse individuals. (Freeman 1999;Johansson 1987) argue that this needs not be the case, and that it depends upon income and price elasticities.

Non-use value refers to the value beyond current or future consumption. For instance, an individual who never sees a polar bear in real life might feel a need to secure its survival; i.e. the individual values the existence of the polar bear. Furthermore, the individual might assign value to the opportunity of the next generation to experience the polar bear - this is termed bequest value. These descriptions of non-use value have been developed by environmental economists and hence tailored for this field. In health care non-use value might rather be characterised as a ‘caring externality for other peoples’ health’ that implies an altruistic way of thinking. Altruistic value can be divided into a concern about the general utility levels of others (i.e. ‘individual altruism’) and a concern about the consumption of specific goods by others (i.e. ‘paternalistic altruism’). For example, giving money to the poor is individual altruism, i.e. “it brings me utility to help others consume a good that gives them utility”. On the other hand, a non-smoker being willing to attend a stop-smoking course is paternalistic altruism, i.e. “I know what is best for you and therefore I decide for you”. And here is the problem: What if smokers do not want to attend the stop-smoking course? Then we put value onto a good where the potential consumers do not want to consume the good, i.e. the good has a negative utility for the consumers. In the worst case we might end up

offering a good that no one wants to consume or, alternatively, end up valuing the good twice – referred to as double counting.

Another non-use value is the value of positive externality. That one individual quits smoking has a positive externality on other individuals, as the cigarette smoke will no longer bother them. In this case the smoker (user) will indirectly affect the non-smokers (non-users) and hence will experience a gain in utility. The exclusion of such non-use value might result in an underestimation of the associated value of a good.

In general the clarification of non-use value can seem problematic and even though economists working in the field of CBA accept the hypothesis of non-use value, there is little consensus as to terminology, definitions, what motivates people to hold non-use values, and how to measure non-use value empirically. Most of the literature that discusses non-use value is from environmental economics (e.g. Carson 1999;Freeman 1999;Hanemann 1994;Hanemann 1995), however the issue is also sporadically discussed in health economic literature, (e.g. O'Brien & Gafni 1996;Olsen 1997). Figure 2.1 gives an overview of the components of total economic value by summarizing the concepts from different sources of literature.

From a health economic perspective, the above description of economic value imply categorizing of individuals into three groups: Currently diseased, currently non-diseased but at future risk and currently non-diseased and not at future personal risk (O'Brien & Gafni 1996;Olsen 1997). In general, the majority of studies in health care have surveyed the first category of respondents: Diseased who are users of the treatment programme.

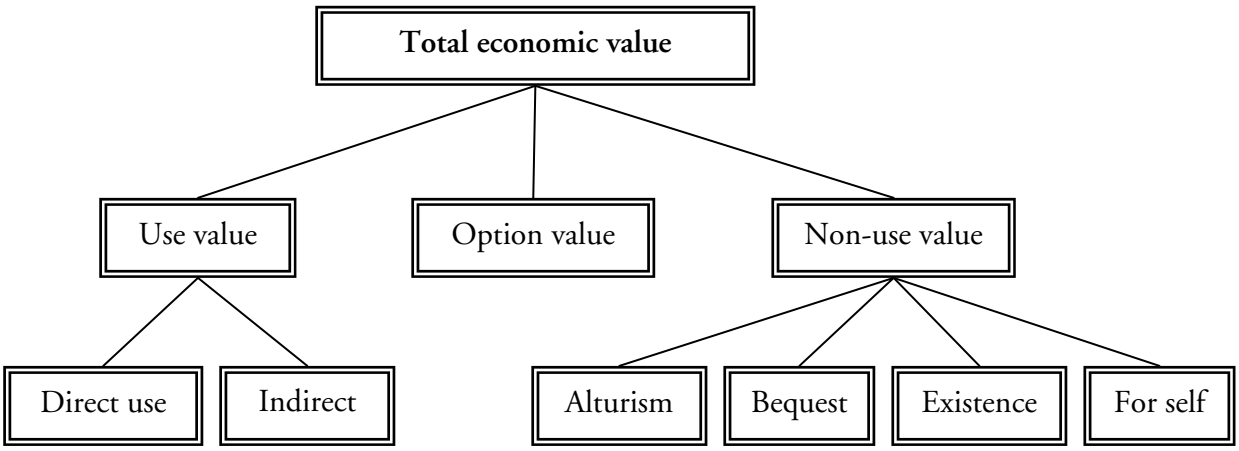

Figure 2.1: Total economic value. (Source: Bateman et al. 2002;Boardman et al. 2001;Freeman 1999;Garrod & Willis 1999)

2.3 Stated versus revealed preference methods

As mentioned at the beginning of this paper, there is a school of thought within health economics that believes that society's preferences for alternative allocations should play a part in determining which health care interventions are provided to the public. A number of methods have been developed that attempt to elicit individuals' preferences for use in priority-setting. Preference-based outcome measurement is in general divided into two approaches: stated preference methods (SP) and revealed preference methods (RP). Depending on the goods in question, both types of methods can be useful. As the name indicates, RP is a generic term for market analysis and refers to the observation of preferences revealed by real market behaviour. A prerequisite for applying RP is that there is a market demand curve for the good in question. Economic evaluation is often more complicated than this, however. The rationale for public polices and the desirability of evaluating them with CBA is that either the market for the good does not exist or the market is imperfect. In relation to economic evaluation, the term 'revealed preference' is often more restricted in the sense that RP in this context refers to analysis of individuals' preference structures for the given good based on preferences for a closely related (complementary) good that is available on the market - hence termed indirect RP valuation. The price of the marketed good is thus used as an indicator of the value of the non-marketed good. For instance, in an analysis of preferences regarding risk reduction for car accidents, it might be useful to perform a market analysis of airbags, where the value for airbags is used to represent the value of risk reduction. Compared to SP methods that are capable of capturing total economic value, RP methods merely capture 'use value'. The

measurement of welfare differs between the RP and SP approaches. Whereas the RP approach relies on the market demand curve, SP data rely on the income-compensated demand curve. The welfare measure in RP studies is the consumer surplus, whereas SP studies estimate true welfare measures – the compensating variation and the equivalent variation. The two demand curves and the associated welfare measures are illustrated in Appendix I.

A problem with RP is that attributes are collinear in market data, making it difficult or impossible to predict the effect of independent variation in an attribute. In such cases, SP data have a key advantage over market data. However, there are important potential problems with the use of SP data as well – SP specific biases. Most obviously, respondents have no incentive to make choices in an SP experiment in the same way they would in the market. Even if people did respond as if they applied their true utility weights to the attributes presented in the experiment, the SP choice situation is typically somewhat different from a market choice situation. Some aspects of the market choice context, such as search costs, are absent in the SP experiments; moreover, the experimental alternatives are defined solely by the attributes presented, while in market data there may be attributes observed (or perceived) by the consumer but unobserved by the researcher, and these will be pushed into choice model error terms (Keane 1997). Table 2.1 provides a general overview of the advantages and disadvantages of the RP and SP methods

A new area of data modelling involves the combining of RP and SP data. This allows the advantages of each method to be retained while mitigating their limitations. Such data combination appears to improve data quality.

Indirect RP methods include approaches such as the *hedonic pricing* and the *travel cost method*, whereas stated preference methods include *contingent valuation* (CVM) and *discrete choice experiment* (DCE); see Figure 2.2. A short description of each of these valuation methods is given in the next sections.

Table 2.1: An overview of the revealed preference and stated preference methods

	Revealed preference **- Observed behaviour**	**Stated preference** **- Hypothetical behaviour**
Approach	Consumers' preferences are revealed through their actions in real markets which are related to the value of interest	Consumers are asked to state their preferences for hypothetical scenarios/alternatives that comprise a set of attributes and different levels of these attributes
Direct methods	Competitive market price (observation of market prices)	Contingent valuation method (directly asking individuals their WTP)
Indirect methods	Travel cost method Hedonic pricing method (observation of choices in referendum settings)	Discrete choice experiment (estimation of WTP by use of price variable)
Applicable goods	Real goods	Hypothetical and real goods
Demand curve	Marshallian demand curve (observable demand curve in market analysis)	Hicksian demand curve (compensated demand curve, adjusted for income)
Welfare measure[5]	Consumer surplus (not a true measure)	Compensating variation – CVM (and surplus[6]) Equivalent variation – EV (and surplus)
Disadvantages	• Limited to the supplying of information regarding values that have been experienced • Limited number of cases where non-market values/goods exhibit a quantifiable relationship with market goods • Choice sets, attributes of choice options and individual characteristics are NOT controlled and/or designed a priori but rather occur/co-occur	• Observed preferences may not reflect actual behaviour • Absence of incentive for the respondent to provide accurate responses • Incentive for the respondent to behave strategically • Costly evaluation • Vulnerable to violation of economic decision-making
Advantages	• External validity is maximized because the choices observed are real market choices in which consumers have committed money, time and/or other resources • Low-cost evaluation	• Provides preferences and information that are otherwise impossible to reveal when actual choice behaviour is restricted in some way • Allows the researcher complete control over the choices offered and their attributes • Ensures sufficient variation in data

Sources: (Boardman et al. 2001;Garrod & Willis 1999;Hanley & Spach 1993;Hidona 2002;Train 1993)

[5] (Willig 1976) outlined conditions under which consumer surplus and the Hicksian welfare measures approximate each other: where income effect is small (income elasticities are not significantly different from zero) and substitutions are available.

[6] Surplus measures are characterised by a quantity constraint on the good in question, and hence restrict the individual to consuming a specific quantity of the good whose price/quality has changed (Freeman 1999).

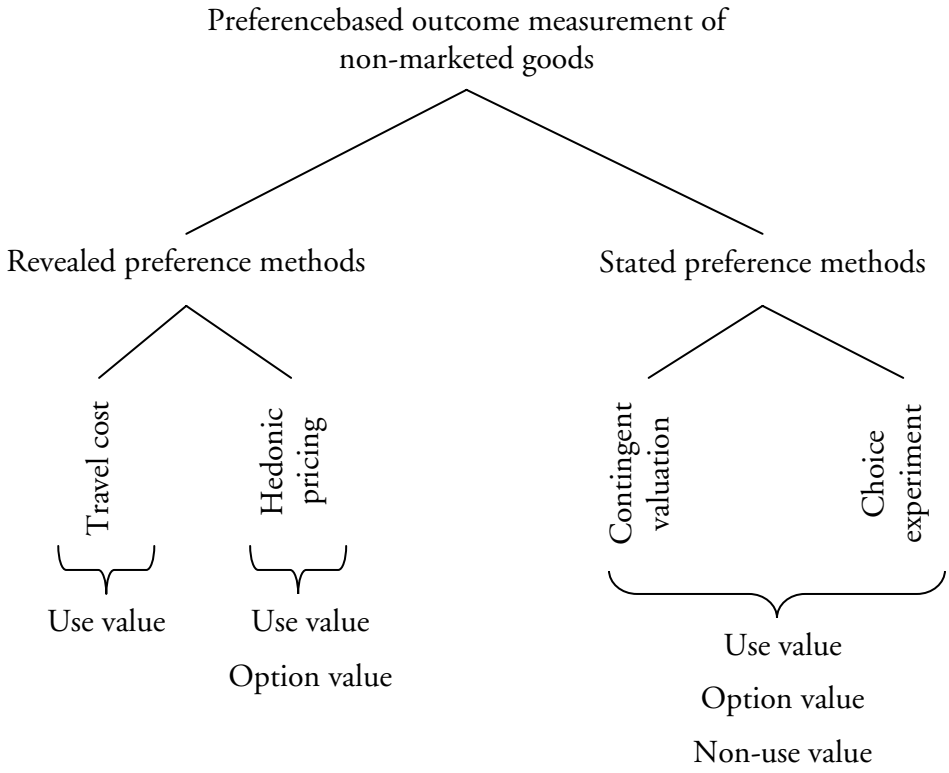

Figure 2.2: Valuation methods and their relation to economic value

2.3.1 The hedonic pricing method

The first environmental studies using the hedonic pricing method were published in the late 1960s and early 1970s (Hanley & Spach 1993). The hedonic pricing approach is a method of ascertaining the value of or the pleasure felt from attributes of a good. Although the history of hedonic pricing goes back to the late 1920s, it was not until 1974 that the theoretical foundation for hedonic pricing was developed (Rosen 1974). The hedonic approach – in common with the discrete choice experiment approach – is consistent with Lancaster's theory of characteristics of a good (Lancaster 1966), implying that the method regards a good as a set of attributes and considers the value of a good as a function of each attribute of that good. The value of an attribute is called an implicit price (a hedonic price) of the attribute, because it cannot be observed in a real market. The researcher can estimate the price, however, by analyzing the prices of a good that has different quantities of each attribute in the market. This is the fundamental setting of the method. The hedonic pricing is defined as a method of finding out these implicit prices. The function that determines the market price of a good based on these attributes, is called the hedonic price function (Hidona 2002).

Hedonic pricing has its origin in Labour and Property and different varieties of the hedonic approach can be found. One is the hedonic wage method that has been applied empirically to measure the value of risk reduction (wage-risk trade-off) and the value of local environmental, cultural and social amenities (wage-amenities trade-off) (Freeman 1999). The most common application of hedonic pricing in environmental valuation is in relation to the public's willingness-to-pay for housing (e.g. Garrod & Willis 1992;Tyrvainen & Miettinen 2000). In this context, each property is assumed to constitute a distinct combination of attributes that determines the price which a potential purchaser or tenant is willing to pay. Willingness-to-pay depends upon the existence and level of a wide range of housing attributes that include:

- Structural characteristics, e.g. number of rooms, plot size
- Local characteristics, e.g. quality of nearby schools, local taxes
- Local amenities, e.g. environmental aspects, access to services

By holding all housing characteristics constant except for one (and comparing this with the differences in prices), it becomes possible to estimate a demand function and the marginal value of the characteristic in question - for instance the value of living in close proximity to a forest. Basic problems with the hedonic pricing method include omitted variable bias, multicollinearity, choice of functional form, market segmentation and restrictive market assumptions.

2.3.2 The travel cost method

The travel cost method is the oldest non-market valuation technique and was developed for use in environmental valuation. The logic behind the method was first described by Harry Hotelling in 1947 in a letter to the US Parks Service and was subsequently developed, among others, by Clawson and Knetsch (Hanley & Spach 1993). The travel cost method has most often been applied to value recreational demands such as the value for fishing, hunting, boating and forest visits (e.g. Brainard et al. 2001;Hesseln et al. 2003;Willis & Garrod 1991). Like hedonic pricing, the travel cost method seeks to place a value on non-market goods by using consumption behaviour in a related market, where travel costs are used as a measure of the preferences for the good. The rationale behind the travel cost method is that as the price of access (i.e. cost of travel) increases, the visit rate tends to fall. By examining these two parameters is becomes possible to estimate a

demand curve and hence consumer surplus (Garrod & Willis 1999). Basic problems with the travel cost method include choice of dependent variable, multi-purpose trips, holiday-makers versus residents, calculation of distance costs, and the value of time and statistical problems.

2.3.3 The use of revealed preference methods in health economics

The hedonic pricing and travel cost methods have rarely been used in the valuation of health care. This is mainly due to two reasons: firstly, the limited use of preference-based economic evaluation in general and, secondly, the limited applicability of the methods. (Delucchi et al. 2002) attempted to evaluate the disutility from air pollution (health effect and visibility) by means of hedonic house pricing. Schumacher & Whitehead (2000) used the hedonic wage model to examine the marginal value of medical care inputs such as the physician-to-population ratio and the availability of specialist services. The results indicated a relationship between an increase in the number of physicians per 10,000 population and lower wages for underserved regions. The travel cost method may be useful in the evaluation of health care interventions in which travel (and hence travel costs) plays a vital role, such as an evaluation of the benefits of living close to a casualty department. Clarke (1998; 2002) used the travel cost method to estimate the benefits of a mobile mammographic screening unit in rural areas of Australia and found that the benefits depended on the distance to the nearest fixed mammographic screening unit – more precisely, the benefits of mobile mammographic screening exceeded the costs if a rural town was situated 29 km or more from a fixed mammographic screening unit.

2.4 The contingent valuation method

The contingent valuation method (CVM) and discrete choice experiment (DCE) are two types of SP methods. SP methods are the only methods in which the total economic value can be measured, as they can incorporate non-use value and option value, thus making it possible to value hypothetical goods and interventions. This characteristic has far-reaching potential as it implies that SP can be used to value potential future goods and interventions and hence to be used as a tool to guide research and development in a direction beneficial to society.

CVM was developed in the USA in the 1960s and is well represented in the literature – especially in the field of environmental economics. CVM attempts to measure the value of a good holistically, i.e. valuing the good in its entirety, by asking people directly about their willingness-to-

pay (or willingness-to-accept). In CVM nothing is revealed about the value of the different attributes that comprise the good (in comparison to the discrete choice experiment, which estimates the values of the attributes that make up the good). This is an important distinction between CVM and DCE and is the reason why many researchers consider DCE to possess several advantages[7] (Bateman et al. 2002;Bennett & Blamey 2001;Hanley et al. 1998). In 1989 the US experienced an environmental catastrophe that influenced the development and future use of CVM and SP methods in general: The Exxon Valdez oil spill. The Exxon Valdez oil spill was the biggest in US history and forced authorities to change the laws and regulations dramatically. Ten million seabirds, hundreds of otters, porpoises, sea lions and several species of whale were under severe threat. While Exxon Valdez was considered liable for payment of cleaning and damage costs, an extensive debate – which later resulted in a lawsuit – was initiated concerning whether or not Exxon Valdez should also be charged for the loss of biodiversity. It was argued that the public experienced a decline in utility due to the loss of biodiversity and that this loss should be compensated. This debate led the US National Oceanographic and Atmospheric Administration (NOAA) to appoint a panel to report on the potential use of CVM in the valuation of non-marketed goods, such as biodiversity, and whether such methods could be considered as an acceptable basis for pursuing court cases (Arrow et al. 1993). Although the lawsuit against Exxon Valdez was never completed due to an agreement among parties, the NOAA panel legitimized the use of CVM by concluding that it could be the basis for estimation of non-use values for inclusion in natural resource damage assessment cases (Boardman et al. 2001).

Depending on its design, CVM can be categorized as Bateman et al. (2002):

- **Open-ended**: An individual is asked to state his/her maximum willingness-to-pay; no amounts are given beforehand. This approach was the first to be used, but has been subject to much criticism, for example, due to the possibility of obtaining unrealistic responses.

- **Bidding game**: In this approach respondents are faced with several rounds of discrete choice questions. The bidding game is continued until the respondent expresses unwillingness to pay the given amount. This was one of the most common methods in use; it

[7] The advantages and disadvantages of CVM and DCE are discussed more fully in chapter 6

is rarely used today, however, due to considerable evidence of strong biases such as starting point bias.

- **Payment card**: The individual is confronted with a given set of amounts and has to identify the most preferred amount. This approach was developed as an improved alternative to the open-ended and bidding game formats.

- **Dichotomous choice CVM** (also termed **close ended** and **referendum CVM)**: An individual is confronted with an amount and has the opportunity to accept or reject to pay the given amount. Studies have indicated that dichotomous choice results in higher WTP estimates compared to open-ended and payment card approaches. This elicitation format is thought to simplify the cognitive task faced by respondents. The NOAA panel recommends the use of dichotomous choice CVM.

- **Double bounded dichotomous choice**: The dichotomous choice question is followed up by another dichotomous choice question depending on the prior answer. This format has gained terrain, as it is more efficient than its counterpart - more information is gathered from each respondent.

The earliest CVM studies used simple open-ended elicitation. During the 1980s a number of authors developed the framework for dichotomous choice CVM (e.g. Hanemann 1984). It was argued that dichotomous choice (and the payment card approach) was less susceptible than open-ended CVM to incentives of over- and understating true WTP. The NOAA panel also recommended the use of dichotomous choice, as the only methodologically acceptable elicitation format. During the 1990s, criticism of the dichotomous choice approach began to appear, based on the large (and hence costly) sample size requirement due to the low amount of information gathered from each respondent, and the limitations of analysing goods with multidimensional changes. For such reasons, researchers (dominated by environmental economists) started to look for other methods for the evaluation of non-marketed goods, and turned towards the DCE (Foster & Mourato 2003). The shift towards DCE seemed very obvious as the dichotomous choice and DCE share a common theoretical foundation of random utility theory (which is discussed more fully in chapter 4).

2.4.1 CVM in health economics

CVM was the first SP method to be used for valuation in health care, starting in the late 1980s. The article by O'Brien & Gafni (1996) is one of the first of its type in health economics, discussing the potential use and advantages of CBA in health care. As mentioned earlier, CVM had the odds against it due to the fact that health economics, in comparison to environmental economics, had a long history of using cost-effectiveness analysis (CEA), and later cost-utility analysis (CUA), as an evaluation tool. The conditions for expansion of the CVM approach were therefore everything but perfect. Nevertheless, CVM is now recognized as an alternative to standard effect measurement – especially by health economists who question the foundation of CEA/CUA and who believe that CVM can make a positive contribution in the evaluation of health care interventions. (For further discussions and review of CVM in health care see, for example, (Diener et al. 1998;O'Brien & Gafni 1996;Olsen 1997;Ryan et al. 2004;Shiell & Gold 2003;Smith 2003).)

The other main stated preference method is the discrete choice experiment. This is the focus of the remainder of this paper.

3 Fundamentals of the discrete choice experiment

For many years discrete choice experiment (DCE) and similar techniques have been of interest to researchers in a variety of academic disciplines. Probably for that reason there has been (and still is) little consensus on the content, naming and theoretical foundation of the methods. As stated by Garrod & Willis (1999):

> *"Choice experiments may be found in the literature under a variety of guises, and confusion may arise from the different terms which are used to describe the various techniques which fall into this category"* (Garrod & Willis 1999, pp 203).

Choice techniques have been used by psychologists since the 1960s (e.g. Anderson 1962; Luce & Tukey 1964) and in the early 1970s were introduced to the marketing literature, where they received much attention from both academic and industrial circles (e.g. Green et al. 1972;Green & Rao 1971). In the marketing field, the techniques became known as *conjoint analysis*[8], a term conceived by Green and Srinivasan (1978). Conjoint analysis has played an important role in the prediction and understanding of consumers' decision-making and choice behaviour. During the late 1970s and the 1980s the development and application of the conjoint analysis approach increased dramatically. Wittink & Cattin (1989) estimated that 400 marketing studies using conjoint analysis were carried out per year during the early 1980s.

Contemporary to the development and application of conjoint analysis, the economic literature (especially in the transportation field) reported new ways of modelling discrete choices (disaggregated models) and the theoretical foundation for the modelling was developed using a theory known today as random utility theory (see Ben-Akiva & Lerman 1985). The development of random utility theory and disaggregated models became the benchmark for the use of choice techniques in economic literature as it provided the necessary link between observed consumer behaviour and economic theory. Random utility theory provides a comprehensive way to conceptualize and model market behaviour. To specify that these choice approaches are founded in economic theory (compared to the methods used in marketing), the term 'conjoint analysis' is no longer widely used in the economic literature. Louviere argues that the term 'conjoint analysis' should be replaced with a more appropriate term so as to indicate that the techniques are based on

[8] The term conjoint analysis originates from the merging of the terms 'considered' and 'jointly'

random utility theory (Louviere et al. 2000;e.g. Louviere 2000;Louviere 2001a)[9]. Ryan & Wordsworth (2000) note that the choice not to use the term conjoint analysis helps to distinguish choice-based experiments used in economics from other forms of conjoint analysis that do not derive from economic theory. In the environmental economics literature, choice techniques that are founded in economic theory are grouped under the terms 'choice modelling' or 'choice experiments'. There is now a general consensus to divide the choice techniques into three categories to reflect differences with respect to theoretical assumptions, methods of analysis and experimental procedures (Bateman et al. 2002;Blamey et al. 2002;Louviere et al. 2000)[10]:

- Discrete choice experiment[11]
- Contingent ranking
- Contingent rating

These three techniques have much in common (see Box 3.1 for an illustrative example of each choice technique). The basic designing of the alternatives is the same in each approach and the respondents must decide which of mutually exclusive multi-attribute alternatives they prefer. Furthermore, all three techniques – under the right assumptions - can be shown to be consistent with welfare economic theory. In DCE the respondents have to choose one alternative out of a given number of alternatives (two or more). As the DCE only contains information regarding the preferred alternative, the data can be said to be weakly ordered. Contingent ranking, in contrast, requires all the alternatives to be ranked, and the data therefore provide a complete preference order (strongly ordered). While a contingent ranking exercise contains more information about preferences than a similar discrete choice exercise, it is also more cognitively demanding. The degree of task complexity in contingent rating is even higher as the respondents have to place a value (characterising the strength or degree of preference) on each alternative (Louviere et al. 2000). Contingent rating (compared to contingent ranking and DCE) provides the respondent with the opportunity to rate alternatives equally (named ties) and thus to indicate indifference between

[9] "*Thus, the bottom-line is that conjoint analysis is what conjoint analysis researchers do. Unlike theory and methods in the hard sciences, the original conjoint analysis theory is not a theory about the behaviour of preferences or choices. It is a theory about the behaviour of sets of numbers in response to factorial manipulations of attributes*" (Louviere 2000, pp 3)

[10] Bateman et al (2002) define a fourth type of procedure named 'paired comparison'. This method is also known from the contingent valuation literature and is closely related to the contingent rating approach. For an example of the application of the paired comparison approach, see (Johnson et al. 2000)

[11] Also known as choice experiment, and (discrete) choice analysis

alternatives. The modelling of ranking and rating data differs slightly from that of DCE data due to the stronger ordering of alternatives. Models used for ranking and rating data include the rank-ordered logit and ordered probit. The use of ranking and especially rating techniques suffers from potential theoretical and practical obstacles. These concerns include the difficulty individuals might experience ranking/rating all the alternatives; and the fact that rating tasks in particular involve difficulty in making interpersonal comparisons and departure from the choice contexts that are faced by consumers in the real world (Bennett & Blamey 2001). Bateman et al. (2002) also argue that the methods differ in their ability to produce WTP estimates that can be shown to be consistent with the usual measures of welfare change, and which thus can be used as part of a CBA. Today the DCE is the most applied choice modelling approach in the economic literature, whereas contingent rating is hardly ever used. For an empirical comparison of the three choice approaches refer to (Boyle et al. 2001).

The DCE is the simplest of the choice techniques and thus its biggest advantage is the low cognitive complexity – the degree of task complexity and difficulty arising from the experiment (Louviere et al. 2000). As mentioned earlier, the DCE is closely related to dichotomous choice CVM, as both methods involve consumers making mutually exclusive choices from a set of substitutable goods. The methods also share the same economic foundation, random utility theory. The DCE approach in the form as it is known today was developed in the early 1980s, with Louviere & Woodworth (1983) being the first to use the term 'choice experiment' (Hanley et al. 2002).

While CVM was applied earlier and more widely in environmental economics than health economics, choice techniques were introduced to the environmental and health economics literature at a similar time in the early 1990s (Hanley et al. 2003;Hundley et al. 2001;Johnson et al. 2000)[12]. DCE has had limited use in health economics compared to environmental economics, however. The reasons for this are the strong tradition for application of preference-based methods in environmental economics and the limited interest shown by the pharmaceutical industry and governments in the application of preference-based economic evaluation in general. This has slowly started to change, however, and within the last five years there has been growing interest for DCE within health economics with respect to both applications and methodology. The UK[13] is now the

[12] In 1994 the DCE was introduced to environmental management by (Adamowicz et al. 1994;found in Hanley et al. 1998).

[13] (e.g. Bryan et al. 1998;Cairns & van der Pol 1997;Farrar et al. 2000;McIntosh & Ryan 2002;Ratcliffe & Buxton 1999;Ratcliffe & Longworth 2002;Ryan 1994;Ryan 2004a;San Miguel et al. 2002;Scott & Vick 1999;Ubach et al. 2003)

leading contributor of DCE papers, while countries such as Australia[14] and USA[15] also play an important role in the dissemination of the DCE approach. In Denmark and other Nordic countries there has also been a growing interest in DCE (e.g. Bech 2003;Carlsson & Martinsson 2003;Gyrd-Hansen & Slothuus 2002;Skjoldborg & Gyrd-Hansen 2003). This increased interest in the DCE may be a reflection of the growing recognition that DCE offers many different possibilities in the context of decision-making, when compared to the use of standard evaluations such as cost-effectiveness analysis. Government reports in England have recommended the involvement of consumers in the planning of health services, indicating a change to encompass preference-based methods (Hundley et al. 2001). For the latest literature review of the application of DCEs (from 1990-2000) in health care, see Ryan & Gerard (2003); for a well-written example of the application of DCE in health care, see Scott & Vick (1999) who used DCE to examine the patient–doctor relationship in a principal–agent theoretical foundation.

3.1 The discrete choice experiment

The term 'discrete choice' arose from the distinction between continuous and discrete variables for denoting the set of alternatives. The word 'discrete' indicates that the choice is discrete in its nature, meaning that it is only possible to choose one alternative. A discrete choice situation is defined as one in which the respondent faces a choice among a set of alternatives meeting the following criteria (Train 1993):

- The number of alternatives in the set is finite
- The alternatives are mutually exclusive
- The set of alternatives is exhaustive (all possible alternatives are included)

While most CVM questionnaires ask respondents only one question regarding a 'proposed' situation, the DCE often asks respondents to make a sequence of choices. The DCE is characterized as a method in which the good in question is described by a number of attributes. Choice experiments can thus be used to examine the response of the individual to changes in the scenario attributes. Rather than examine the entire scenario as a package, the choice experiment allows the researcher to break down the relevant attributes of the situation and to determine preferences for

[14] (e.g.Hall et al. 2002;Hall et al. 2003;Jan et al. 2000;Salkeld et al. 2000;Salkeld et al. 2003;Taylor & Armour 2003)
[15] (e.g. Johnson et al. 1998;Johnson et al. 2000;Maddala et al. 2003;Phillips et al. 2002)

different attributes (Garrod & Willis 1999). The proposed alternatives in each choice are all different in terms of the state of the good described to the respondents. These descriptions are known as attributes of the alternatives. The variation across the alternatives in the choice sets is achieved by assigning different levels to the attributes, according to a systematic process known as experimental design (this is described more fully in chapter 5).

The attributes and their levels must be constructed so that they force the respondent to trade[16]. For instance, a respondent may rather drive a car to work than take the bus or train, but the lower price associated with public transportation makes him choose differently (see Figure 3.1). It is important to note that each time a level is changed, a new scenario (i.e. a different package of the good) arises. By securing a certain variation in the scenarios, it becomes possible to examine the degree to which each attribute influences the choice of the decision-maker; that is, to estimate the marginal rates of substitutions of the attributes (Louviere et al. 2000). The specific conditions to be fulfilled are discussed later in this chapter.

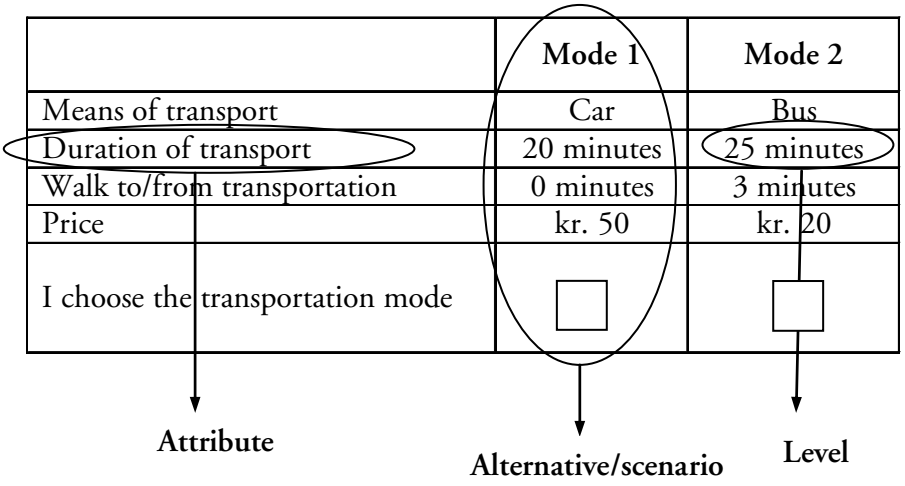

Figure 3.1: A choice set

Attributes can possess either positive or negative utility, and to varying degrees (depending on the level of the attribute). Attributes related to a health care intervention (e.g. opportunity for

[16] The theoretical foundation of the DCE is built upon the neoclassical consumer theory (the compensatory decision-making principle). See chapter 4 for more information.

treatment and health effects) are expected to possess positive utility, where utility increases as the effect increases, i.e. people value an effective treatment higher than a less effective treatment. On the other hand, process utility and side effects are expected to possess negative utility, i.e. people want the treatment process to be as comfortable as possible and they want to avoid side-effects. Normally, it is a good idea to ensure that the DCE (and hence the alternatives) possesses both positive and negative utilities as this will improve the likelihood that respondents will trade-off one attribute for another; this is not a requirement, however. An advantage of quantitative attributes is that their levels can be assumed to be linear in utility, whereas qualitative attributes have to be divided into dummy variables. Two attributes in particular have some special features that make them very appropriate for use in DCE: these are time and cost. Both are characterized as quantitative, constrained and negatively valued attributes, producing time-trade-offs[17] and marginal WTP estimates, respectively. Despite the fact that essential issues – theoretical as well as methodological – are related to the inclusion of these attributes (in particular the cost attribute) such issues have only received little attention in the (health) economic literature.

3.1.1 Inclusion of a cost attribute

The cost attribute plays an important and distinct role in the DCE. The inclusion of a cost attribute provides the DCE with a special quality as it becomes an elicitation procedure for willingness-to-pay (WTP). This implies that benefits are estimated in monetary terms and causes the DCE to be consistent with welfare economics (i.e. the potential Pareto improvement condition). Results from different studies can then be compared and - on the grounds of economic efficiency - used in priority-setting. Inclusion of a cost attribute makes it possible to indirectly obtain the respondent's WTP for either the good in its entirety (an alternative) or the respondent's WTP for the attribute respectively, i.e. marginal WTP (also termed part worth or implicit price) (Bennett & Blamey 2001). The method is indirect in the sense that respondents are not directly asked their WTP as in the CVM methods (especially open ended), but instead have to trade cost for improvements in the positively valued attribute (or for a decrease in negatively valued attributes). Marginal WTP is simply the marginal rate of substitution in which the numeraire is the cost attribute. The estimation of welfare is based on using the coefficient of the cost attribute as a proxy for marginal utility of income (money). This indirect approach of estimating WTP is often

[17] (e.g. van der Pol & Cairns 2001)

considered to be an advantage over CVM as it considerably reduces focus on the price aspect (Blamey et al. 2000).

Costs can take many different forms in a DCE, including options such as consumer price, transportation cost, salary, donation, tax payment, tax payment in a referendum context, etc. The form in which cost (payment) is specified in the survey, the conditions under which it is required and the link between response and potential payment is termed the 'payment vehicle' (Green et al. 1998b). The choice of payment vehicle depends on the context of the choice task and the choice condition. This is discussed more fully in section 5.3.

Box 3.1. Examples of choice techniques

Imagine a situation where the mode of transportation to work has to be decided.

1. The discrete choice: (choice of one option from a set of competing options)
Choose which of the following transportation modes you prefer the most (check one box only).

	Mode 1	**Mode 2**	**Mode 3**
Means of transport	Car	Bus	Train
Duration of transport	20 minutes	25 minutes	15 minutes
Walk to/from transportation	0 minutes	3 minutes	10 minutes
Price	kr. 50	kr. 20	kr. 30
I choose the transportation mode	☐	**X**	☐

The information that is obtained using this DCE:
Transportation mode 2 > Transportation mode 1 and Transportation mode 3

2. Contingent raking: (a complete ranking of options from most to least preferred)
Rank (A, B and C) the following transportation modes such that 'A' is the transportation mode you most prefer, 'B' is a better transportation mode than 'C' but is worse than 'A', and 'C' is the transportation mode you least prefer.

	Mode 1	**Mode 2**	**Mode 3**
Means of transport	Car	Bus	Train
Duration of transport	20 minutes	25 minutes	15 minutes
Walk to/from transportation	0 minutes	3 minutes	10 minutes
Price	kr. 50	kr. 20	kr. 30
I rank the transportation modes	**B**	**A**	**C**

The information that is obtained using this contingent ranking:
Transportation mode 2 > Transportation mode 1 > Transportation mode 3

3. Rating: (expressing degrees of preference by rating options on a scale)
Rate the transportation modes from 1 to 10, where '1' is the worst transportation mode you can imagine and '10' is the best.

	Mode 1	**Mode 2**	**Mode 3**
Means of transport	Car	Bus	Train
Duration of transport	20 minutes	25 minutes	15 minutes
Walk to/from transportation	0 minutes	3 minutes	10 minutes
Price	kr. 50	kr. 20	kr. 30
I rate the transportation modes	**7**	**8**	**3**

The information obtained using this rating:
Alternative 2 is preferred by 1/10 to Alternative 1 and 5/10 to Alternative 3

4 Foundation of the discrete choice experiment

The theoretical foundation of the DCE is rather complex as it combines several different economic theories. The DCE is based on probabilistic choice theory and named random utility theory and is consistent with Lancaster's economic theory of value and neoclassical economics (Lancaster 1966;Manski 1977). Random utility theory allows the researcher to elicit preferences for complex multidimensional goods, from which models of preferences can be estimated (Hall et al. 2003).

The basis of probabilistic choice theory and modelling is that there is some uncertainty surrounding an individual's choices – we cannot perfectly predict individual choices. An important characteristic of models dealing with uncertainty is that, instead of identifying one alternative as the chosen option, they assign to each alternative a *probability* to be chosen. Many types of discrete choice models have been used in a variety of research areas such as biology, psychology and economics. All probabilistic choice models are characterized by the following equation (for a given alternative *i*),

(4.1) $$U_i = V_i + \varepsilon_i$$

What differs between the models is the way in the variables (terms) are interpreted. Probabilistic choice modelling can be divided into two main families:

(1) The decision rule is assumed to be random and the utility deterministic
(2) The decision rule is assumed to be deterministic and the utility random

The difference between these two interpretations of probabilistic modelling is very important, as it concerns the factors that determine the probability. The first family of models views the individual's behaviour as **intrinsically probabilistic**, which implies that individual behaviour can change according to internal and external factors. The second family of models views the probability as the **inability of the researcher** to precisely formulate individual behaviour. For an overview of the different families of probabilistic choice theory, see Figure 4.1. In the following, the two interpretations of probabilistic modelling will be discussed in further details.

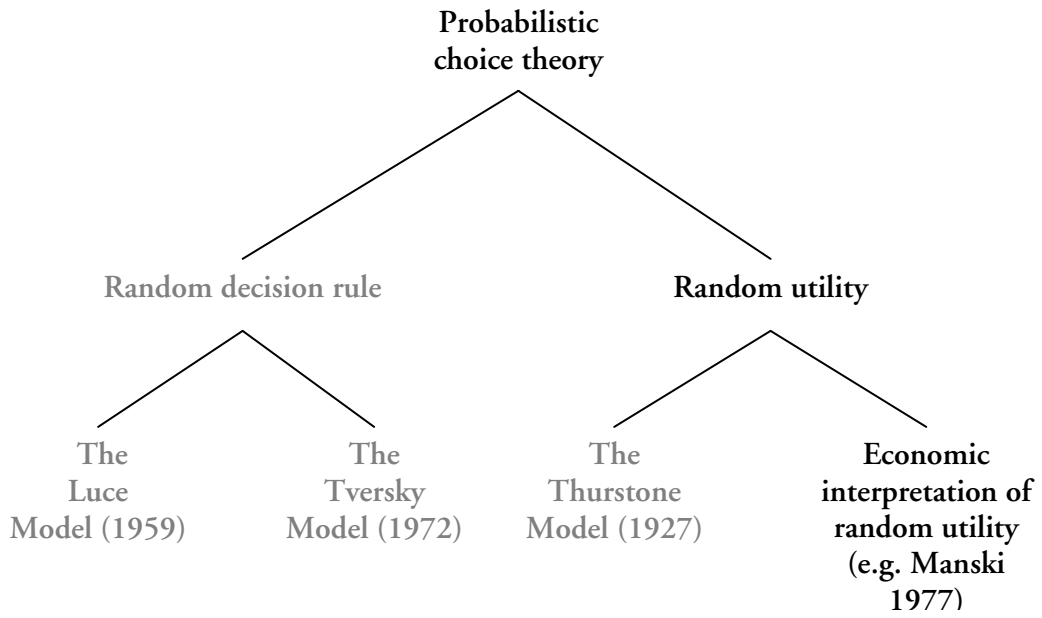

Figure 4.1: Branches of probabilistic choice theory (also termed discrete choice theory). Idea from (Anderson et al. 1991)

4.1 Random decision rule

Probabilistic choice modelling with a random decision rule includes the axiomatic theory of Luce (1959) and Tversky's elimination of aspect theory (Tversky 1972). This approach assumes that the utility (V_i) of the alternatives is fixed (deterministic). Instead of selecting the alternative with the highest utility, the individual is assumed to behave with choice probabilities defined by a probability distribution function over the alternatives that includes the utilities as parameters (Ben-Akiva & Lerman 1985). This implies that the individual does not necessarily choose the alternative that yields the highest level of utility, but instead has a probability of choosing each alternative. As noted by Tversky (1972), people often experience uncertainty and inconsistency when having to choose among alternatives; individuals are often not sure which alternatives to select, nor do they always take the same choice under seemingly identical conditions.

The common use of the term 'rational behaviour' is based on the beliefs of an observer about what the outcome of a decision should be and stands in contrast to impulsiveness, in which individuals respond to choice situations in different ways depending on their variable psychological state at the time a decision is made (Ben-Akiva & Lerman 1985). Economists and psychologists have radically different views of the decision-making (DM) process. The primary focus of the psychologist is to understand the nature of the decision elements, whereas the economist focuses

primarily on the mapping from information inputs to choice, based on rational and utility-maximizing behaviour. Psychological views of the DM process are dominated by ideas that behaviour is local, adaptive, learned, dependent on context, mutable and influenced by complex interactions of perceptions, motives, attitudes and affect. Economics, on the other hand, treats preferences (values) as primitive of the analysis and the decision process as a black box. Psychology has various theories and techniques for studying the DM process. The leading paradigm has been the research of Tversky and Kahneman on experimental study of cognitive anomalies (also called the heuristics and biases approach): circumstances in which an individual exhibits surprising departures from rationality (Ben-Akiva et al. 1999;Gilovich et al. 2002;McFadden 1999).

What is it, then, that constitutes the randomness in utility? The researcher knows neither all the factors influencing the choice nor the precise utility function; furthermore, it is likely that preference strengths vary across individuals (Train 2003). These explanations of consumers' choices are in accordance with neoclassic theory (assumptions of a rational utility-maximising individual), but explanations also exist beyond the frames of standard economic theory. The consumers might not be fully rational in their choice, i.e. they may not have a complete preference function. Herbert Simon proposed the criterion for performance that links economics with psychology, by distinguishing 'perfect' rationality from 'non-perfect' rationality, named *bounded rationality*. Bounded rationality implies that people reason and choose rationally, but only within the constraints imposed by their limited search (cost of information) and computational capacities (limited cognitive capacity). They do not therefore necessarily end up choosing what is best for them (e.g. Simon 1982). Bounded rationality recognizes the constraints on the decision process that arise from the limitations of human beings as problem-solvers, with limited information-processing capabilities. To be able to cope effectively with these limitations, Simon presented the simplifying heuristics that people could employ, such as fast and fugal heuristics; his approach is in many respects in line with the psychological approach of heuristics and biases (Gilovich et al. 2002).

Current experimental economic literature has begun to recognize explanations of human behaviour beyond those based on traditional theory (see Figure 4.2). As stated by McFadden (2001):

> *"The potentially important role of perceptions, ranging from classical psychophysical perceptions of attributes, through psychological shaping of perceptions to reduce dissonance, to mental accounting for times and costs, remains largely unexplored in empirical research*

> *on economic choice. Finally, the feedback from the empirical study of choice behaviour to the economic theory of the consumer has begun, through behavioural and experimental economics, but is still in its adolescence."* (McFadden 2001, pp 31)

There is no doubt that economists can learn a great deal from the psychology literature that could help to explain some of the phenomena observed in experimental economics that not are consistent with standard economic theory. Little effort has so far been made to apply such knowledge of psychological factors that affect judgments and decision-making to DCE. Of special importance is the fact that the DCE itself raises many concerns when it is considered in the light of evidence of the biases and shortcuts which can affect people's judgements and decision-making (Lloyd 2003). Swait & Adamowicz (2001a) argued that two reasons for the lack of studies exploring human-decision making are, firstly, that the psychological and behavioural decision theory literature has not yet been translated into empirical economic analysis and, secondly, that the data used in economic studies tend to vary. In a study of consumer behaviour, Swait & Adamowicz (2001a) found support for the hypothesis that choice behaviour can be affected by context complexity. They proposed further research to examine whether the degree of complexity induces different choice strategies and hence whether the compensatory models used by economists are adequate. In an additional paper, Swait et al. (2002) discussed how to model 'irrationality' such as context dependence, framing effect, etc. It is evident that, as choice modelling becomes more advanced, the tendency towards exploring and explaining human behaviour will increase.

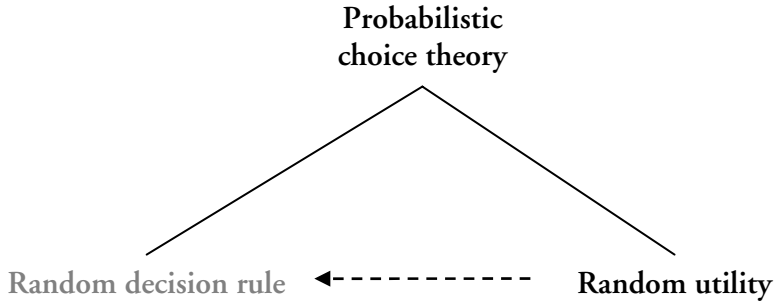

Figure 4.2: Economists have started to recognize explanations of human behaviour beyond those based on traditional theory

4.2 Random Utility Theory

The origins of probabilistic discrete choice modelling goes back to the work of (Thurstone 1927) in psychometrics (*Law of Comparative Judgment*), in which an alternative *i* with true stimulus level is perceived with an error of $V_i + \varepsilon_i$. Thurstone proposed the modelling of individual choice as the outcome of a process in which the random variable is associated with each alternative, and the alternative with the greatest realisation is the one selected (hence belonging to the second family of choice models). When the perceived stimuli are interpreted as levels of satisfaction, or utility, this can be interpreted as a model for economic choice in which the individual chooses the option yielding the greatest realisation of utility (Anderson et al. 1991;McFadden 2001). Marchak introduced Thurstone's work into economics in 1960, by exploring the theoretical implications of choice probabilities for the maximization of utilities that contained random elements (named *Random Utility Models*, RUM). This idea was later taken up and further developed by other economists including Manski and McFadden (e.g. Manski 1977;McFadden 1974).

Consider an individual who has to choose one alternative from a choice set of alternatives. Neoclassical economic theory supposes that the individual has perfect discriminatory power and unlimited information-processing capacity, allowing the individual to rank the alternatives in a well-defined and consistent manner. The individual can thus determine his or her best choice and will repeat this choice under identical circumstances (Anderson et al. 1991). The link with probabilistic choice theory arises from the researcher's lack of information about the individual's true utility function. Probabilistic choice theory is thus introduced not to reflect a lack of rationality in the individual, but to reflect a lack of information regarding the characteristics of the alternatives and/or the characteristics of the individual on the part of the researcher (ε_i) (Manski 1977). The researcher only observes that part of the utility that makes up the alternative. This implies that the utility function is deterministic from the individual's point of view and hence is in accordance with neoclassical economics. The indirect utility function is decomposed into a utility function that depends solely on factors that are observed by the researcher and another utility function that represents all the factors that influence the consumer's choice. Again

$$U_i = V_i + \varepsilon_i \tag{4.1}$$

and, for individual *n*, utility becomes

(4.2) $$U_{in} = V_{in} + \varepsilon_{in}$$

U_i is the true but unobservable (latent) utility for alternative *i*, V_i is the observable systematic component of utility, and ε_i is the factor unobservable to the researcher and treated as a random component (Hanemann 1984). V_i thereby becomes the explainable proportion of the variance in choice and ε_i the non-explainable. RUT assumes that the individual **acts rationally** and chooses the alternative with the highest level of utility - i.e. that the individual is a **utility-maximizer**. As the researcher cannot observe the individual's true utility function, a probabilistic utility function is used in the estimation. The most appropriate probabilistic choice model to apply depends on the assumptions made about the random parameter. This is discussed more fully later in this chapter. Assuming that the individual can choose between two alternatives, *i* and *j*, then the probability that alternative *i* is chosen is given by

(4.3) $$P_i = \text{Prob}(U_i > U_j) = \text{Prob}(V_i + \varepsilon_i > V_j + \varepsilon_j) = \text{Prob}(V_i - V_j > \varepsilon_j - \varepsilon_i) \quad \forall\, i \neq j$$

From this it can be seen that the higher the probability for choosing an alternative, the larger the difference in observed utility. Since probability is defined on a cardinal scale, so are the estimated utility scores (which is the reason why we obtain meaningful WTP estimates). The input of the model is the observed choices, while the output, i.e. what is to be estimated, is the difference in utility for the two alternatives, (V_i–V_j), characterized by the utility for each attribute. Every respondent makes a discrete choice and has chosen either alternative i or alternative *j*. As the choices are aggregated over individuals (taking personal characteristics into account, if possible), the total *observed per cent of the sample* that chooses alternative *i* is interpreted as the *probability* that *an individual* with specific personal characteristics chooses alternative *i*. Thus the choice is transformed to a continuous curve (sigmoid-shape depending on the distributional assumption of the error term, see Figure 4.3) that characterizes the trade-off between the two alternatives. As the quality of the attributes in alternative *i* increases compared to alternative *j*, the probability converges towards 1. This is the same as saying that the probability of choosing alternative *i* increases as the difference in estimated utility between the two alternatives increases. The shape of the sigmoid curve ensures that changes in the utility difference when the individual is very uncertain about which alternative to choose, create large changes in the probabilities; i.e. the model is very sensitive

to changes with probabilities of around 50%. On the other hand, changes in the utility difference have little effect on the overall probability when the individual is more certain of his or her choice. It follows from the above that the probabilities can be interpreted as *preference strengths* for each alternative. Consider a situation in which the probability of choosing alternative *i* is 50%. In this case the utility for each alternative will be the same, i.e. the difference in utility will be zero, and it is impossible to say which of the alternatives individual *n* will choose. No information is contained in this situation as the choice is interpreted as random. In conclusion, the interpretation of the probabilities is what makes it possible to achieve a cardinal utility scale. Such a cardinal scale is necessary to compare the achievable benefits, i.e. to transform the utilities into monetary welfare measures.

The random utility theoretical approach, formalized by Manski (e.g. Manski 1977) and further extended to the modelling framework by McFadden (e.g. McFadden 1974;McFadden 1980;McFadden 1986;McFadden & Train 2000), is in line with neoclassical consumer theory. Manski (1973, fund in Ben-Akiva & Lerman 1985) identifies four distinct sources of randomness:

- **Measurement errors and imperfect information** (i.e. when the data used to estimate the model parameters are not true measures of their theoretical counterparts)
- **Instrumental (or proxy) variables** (the use of closely related variables)
- **Unobserved attributes** (i.e. the choice of alternative is not only determined by the given attributes, but also by some underlying attributes)
- **Unobserved taste variation** (i.e. heterogeneity in preferences). Heterogeneity is a key element to randomness and implies that preferences differ among individuals; the researcher does not know which type of preferences an individual holds and thus cannot perfectly predict the choices made.

Neoclassical theory builds on a number of axioms that give the desired properties to the consumer's preference relation, i.e. they ensure that preferences can be represented by a numerical scale (utility). The axioms ensure that the bundles of goods (alternatives) are ordered by the individual's preference function and that the individual behaves rationally. It is assumed that individuals have complete, stable and consistent preferences and that the indifference curve is continuous (Deaton & Muellbauer 1989). If preferences are not complete, stable and/or consistent then they must be constructed at the time they are elicited, indicating that the process might be

driven by heuristics and affected by context (Swait et al. 2002). The continuity axiom rules out lexicographic orderings, e.g. dominant preferences (although this represents a perfectly reasonable system of choice) and ensures the concept of trade-off[18] which is one of the core principles of the DCE. Violation of one or more of these axioms significantly influences the interpretation of the DCE. Research into axiom violations and their potential reasons (e.g. design issues) is therefore important, as discussed more fully in the next section.

DCE draws upon Lancaster's economic theory of value (Lancaster 1966). This is an extension of the neoclassic consumer theory in that "*goods possess, or give rise to, multiple characteristics in fixed proportions and that it is these characteristics, not goods themselves, on which the consumer's preferences are exercised*" (Lancaster 1966, pp 41). Lancaster's approach regards a unit of any good as a given bundle of attributes of characteristics (for example, a particular type of food will consist of specific flavours, calories, vitamins, etc.) and a combination of goods will produce a vector of quantities of these characteristics. The consumer's preferences are defined over bundles of characteristics and the demand for goods is a derived demand. Consumption is the activity of extracting characteristics from goods. Lancaster's approach is thus very suitable for dealing with the DCE. The amount of an attribute yielded by one unit of a good is fixed, regardless of the level of consumption of this or any other good. Behind this assumption lies the recognition that attributes are objectively measurable and fully known. The inclusion of Lancaster's theory does not violate the neoclassical foundation. Instead of describing the relation between two goods, the MRS describes the relation between two attributes.

4.3 The utility function

In order to analyze a DCE, each respondent's profile has to be collapsed into a single utility number that represents that respondent's overall value. This is done by assigning weights to each of the attributes - after standardizing all values within each attribute to a mean of zero with one unit standard deviation. The weights of each attribute can then be used to derive a linear combination. This approach has the effect of converting each multi-attribute profile into a single point on the real number continuum. Models based on this approach are called compensatory models. The best-known compensatory model is the main-effects additive utility model, in which the weights denote

[18] This is known as compensatory decision-making and is the theory behind the measurement of marginal rates of substitutions (MRS).

the importance assigned to values on each of the attributes. The weights can be viewed as part-worths that make all utility scales commensurate with each other, so that the part-worths (β) can be summed to yield a single (overall) utility. This is the same as saying that the utility of an alternative is equal to the sum of the utilities of its parts (attributes). Treating V_i as a conditional indirect utility function and assuming that utility is linearly additive, the observable utility for alternative *i* can be written as

(4.4) $$V_i = \beta x_i \quad \Rightarrow \quad U_i = \beta x_i + \varepsilon_i$$

where $\boldsymbol{x_i} = (x_{1i}, x_{2i}, \ldots, x_{pi})$ is the vector of the attributes for alternative *i* including a possible price attribute, and β is the weighting (parameters) of the attributes. It is standard practice in a DCE to assume a linear additive utility function (linear in parameter and explanatory variables). Quantitative attributes are coded as they appear, while qualitative attributes are either dummies or effects-coded (Louviere et al. 2000). While the two approaches are functionally equivalent, the use of effects coding facilitates interpretation as the base level impact is made equal to the negative sum of the parameter values for the other categories; dummy coding, on the other hand, incorporates the base level category into the intercept (Mark & Swait 2004).

When comparing (dividing) two attributes, marginal rates of substitution (MRS) are estimated. MRS indicates the trade-off between two attributes that characterize the good and thus the mutual importance of the attributes in question. Holding the overall utility level constant,

(4.5) $$\partial V_i = \beta \partial x_i = 0$$

and MRS becomes

(4.6)[19] $$\mathrm{MRS}_{12} = -\frac{dx_{i1}}{dx_{i2}} = \frac{\beta_1}{\beta_2}$$

[19] In the case of effects coding, it is important to remember that the reference level =-1 and not =0 as in dummy coding. Hence, the coefficient has to be multiplied by 2 in order to obtain the true MRS estimate

When one of the attributes is a cost attribute, the MRS indicates the willingness-to-pay (WTP) for a change in the qualitative attribute, i.e. the marginal willingness-to-pay (MWTP), also known as part-worths. Let the price attribute be denoted as *p.* As income cancels out in linear price models (hence the negative sign of the price variable in equation 4.7), marginal WTP becomes,

$$\text{MWTP}_i = \frac{dx_i}{dU_{income}} = \frac{dx_i}{dp} = \frac{\beta_{x_i}}{-\beta_{price}} \tag{4.7}$$

Applying another functional form for price, such as a log-linear, causes the calculation of MRS to change as the MRS estimates will then be dependent on the absolute level of logged variables. This implies that MWTP depends upon the absolute level of income.

4.3.1 Including interaction terms

It is sometimes beneficial to include interaction terms (variables) in the utility function. Interaction terms can either be interactions between two attribute variables (e.g. $x_1 \times x_2$) or interactions with additional variables, such as person-specific variables (e.g. $s_1 \times x_2$, where *S* denotes the vector of person characteristics). The inclusion of person-specific variables (sociodemographics) in the model makes it possible to account for some of the heterogeneity in preferences between individuals, which can yield very important information, and to perform sub-group analysis. When including interactions, denoted *Z*, the observable utility of individual *n* for alternative *i* is given as

$$V_i = \beta x_i + \beta z_n \tag{4.8}$$

where $z_n = (z_1, z_2, \ldots, z_r)$ denotes the vector of interactions. As person-specific characteristics do not vary over alternatives, they can only enter the model in a way that creates differences in utility over alternatives (recall that the probability and hence estimation of utility is specified as the *difference* in attribute utilities among alternatives) (Train 2003). Person-specific characteristics are thus multiplied by the particular attribute variables which they are assumed to influence. The inclusion of interaction variables necessarily increases the demand for data (number of observations). In particular it is important to note that the inclusion of two way attribute interactions increases the

requirements of the experimental design (see section 5.4.). Moreover, the interaction variables need to be taken into account in the estimation of MRS and welfare measures.

4.4 Discrete choice modelling

In contrast to CVM, modelling is a key element in the application of the DCE. Choice modelling is very closely related to the theoretical framework of RUT, which can make it difficult to separate the two concepts. RUT was developed simultaneously with choice modelling as a way to explain observed behaviour in the setting of economic theory. RUT links observed choices to behavioural assumptions and develops a statistical choice model that explains the observed choices. The overall problem with choice models has been their inherent mathematical and statistical difficulties, especially during the 1970s and 1980s, when much of the model development took place. The availability of faster digital computers over the last 30 years, along with important advances in estimation technology, has made the estimation of choice models easier and the use of more complex models possible (Keane 1997).

There are many models available that can be used in the estimation of the DCE. The choice of model depends, among other things, on the DCE's design. DCEs can be divided into two groups, depending on whether the choice set includes:

- Two alternatives - giving rise to **binary discrete choice** models
- Three or more alternatives - giving rise to **multiple discrete choice** models

From the researcher's point of view, the probability that a given individual chooses an alternative is in the interval between 0 to 1 (0-100 percent). For each choice set, the outcome of the decision is characterized by y. As the observed utility for an alternative increases towards infinity, the probability converges to 1; as the observed utility decreases, the probability converges to 0 (Greene 2003):

$$\lim_{V\to\infty} \Pr(y=1) = 1$$

(4.9)

$$\lim_{V\to-\infty} \Pr(y=0) = 0$$

As mentioned earlier, factors that determine the respondent's choice can be divided into factors that are observed by the researcher (denoted *a*), and factors that are not observed by the researcher (denoted ε). These factors relate to the respondent's choice, as follows:

(4.10) $$y = h(x, \varepsilon)$$

where the function *h[·]* is called the *behavioural process* function. The probability that the respondent chooses a particular outcome, defined as choosing alternative *i*, is the probability that the error term is such that the behavioural process results in that outcome

(4.11) $$P(y \mid x) = \text{prob}\left[\varepsilon \quad \text{s.t. } h(x, \varepsilon) = y\right]$$

To understand this, we define an indicator function $\mathrm{I}\left[h(a, \varepsilon) = y\right]$ that takes the value 1 if the statement of the outcome is true and 0 if false. This is the same as saying that I[·]=1 if the value of ε combined with *a* induces the respondent to choose outcome *y,* and I[·]=0 if otherwise. Then the probability that the respondent chooses *y* is the expected value of this indicator function, where expectations are the possible values of the unobserved factors, i.e. an integral of the indicator for the outcome of the behavioural process

(4.12)
$$P_i = \left(\mathrm{I}\left[\varepsilon \quad \text{s.t. } h(x, \varepsilon) = y\right] f(\varepsilon) d\varepsilon\right)$$
$$\Rightarrow$$
$$P_i = \int \mathrm{I}\left[h(x, \varepsilon) = y\right] f(\varepsilon) d\varepsilon$$

The probability for choosing alternative *i* is given as the integral of the indicator function multiplied by the density function, *f(ε)*, which gives the cumulative distribution function. The probability of choosing alternative *j* therefore becomes

(4.13) $$P_j = P_i - 1$$

To determine the choice probabilities precisely, the distribution of the random variable must be specified, i.e. the integral must be evaluated. Depending on the model being applied, the integral can either take a closed form, a partially closed form or an open form. In the case of an open form integral, the models in question are simulation models; see Table 4.1.

Table. 4.1: Models used to estimate discrete choice experiments.

	Binary discrete choice model	**Multiple discrete choice model**
Number of alternatives in the choice set	Two alternatives	Three or more alternatives
Type of models:		
Complete closed form	Binary Logit	Multinomial logit (MNL) Nested logit (NL)
Partial closed form/ partial simulation		Mixed logit[20], (ML)
Complete simulation	Binary Probit	Multinomial probit (MNP) Heteroscedastic extreme value (HEV)

Sources: (Louviere et al. 2000;Train 2003)

All the models specified in the table have an s-shaped curve (cumulative distribution function) that ranges between 0 and 1; this makes them very suitable for dealing with probabilities (in contrast to a linear probability specification). For an illustration of the s-shaped curve (logit and probit) and a density function (normal), see Figures 4.3 and 4.4.

[20] Also known as the random parameters logit model (RPL)

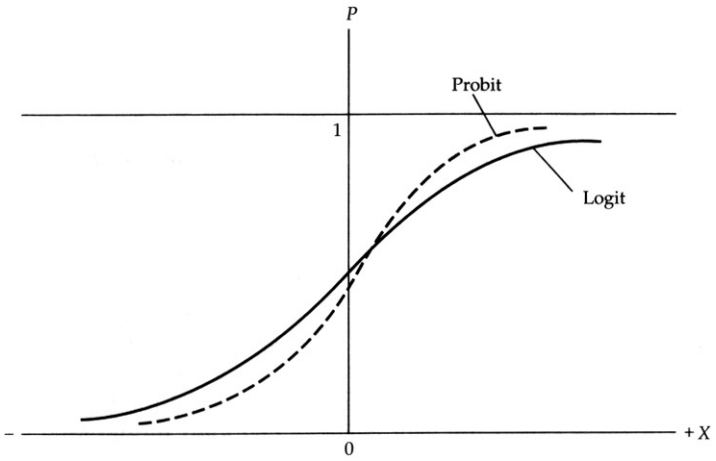

Figure 4.3: Logit and probit cumulative distribution functions

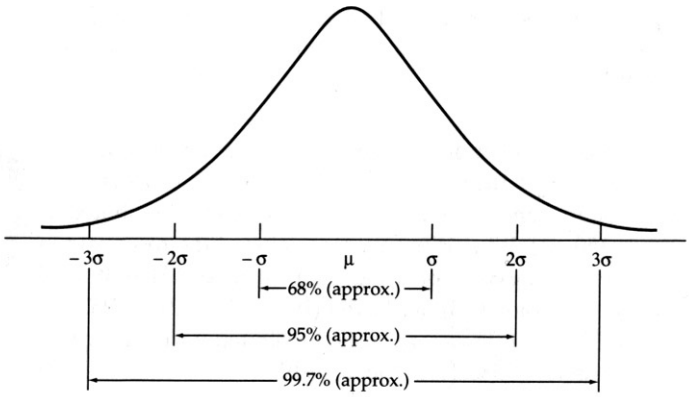

Figure 4.4: Normal density function

Given the joint density function $f(\varepsilon_n)$, the researcher can make probabilistic statements about the respondent's choice. Assuming that alternative *i* differs from alternative *j,* the probability that the respondent chooses alternative *i* can now be written as a cumulative probability function:

$$P_{ni} = \text{Prob}(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj}) = \int_{\varepsilon} \boldsymbol{I}(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj}) f(\varepsilon_n) d\varepsilon_n \tag{4.14}$$

in which $\boldsymbol{I}(\cdot)$ is the indicator function as previously defined. By integrating all the possible values of ε_n , the total probability of choosing alternative *i* is given. Different discrete choice models are

obtained from different specifications of this density, that is, from different assumptions about the distribution of the unobserved portion of utility. The integral can take a closed form, partially closed form or open form depending on the specification of $f(\cdot)$. Logit (binary as well as multinomial) and nested logit have closed form expressions. They are derived under the assumptions that the error terms are independent and identically distributed (iid) and extreme value (i.e. gumbel/logistic) distributed. The probit model (binary as well as multinomial), on the other hand, is derived under the assumption that the unobservable part of utility is normally distributed. Mixed logit is based on the assumption that the unobserved portion of utility consists of a part that follows any distribution specified by the researcher plus a part that is iid extreme value. With probit and mixed logit, the integral does not have a closed form and is evaluated numerically through simulation.

4.4.1 Binary discrete choice models

The binary discrete choice models are the most simplified discrete models and the easiest to apply and interpret. That may explain why they are the most commonly applied discrete choice models. Binary discrete choice models are characterized as models explaining a binary (0/1) dependent discrete variable. The two best-known binary discrete choice models are the binary logit and binary probit models. The difference between the two models is trivial and lies in the weighting of the tails; again see Figure 4.3.

Naming the two alternatives in the choice set, i and j, respectively, the unobserved term, ε, is assumed to be random with the density function $f(\varepsilon)$ where $\varepsilon \equiv \varepsilon_j - \varepsilon_i$. Consider the case where ε is normally distributed, the probability for alternative i can then be solved as

$$
\begin{aligned}
P_i &= \int_{\varepsilon=-\infty}^{V_i - V_j} f(\varepsilon) d\varepsilon \\
&= \int_{\varepsilon=-\infty}^{V_i - V_j} \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2}\left(\frac{\varepsilon}{\sigma} \right)^2 \right] d\varepsilon \qquad (4.15) \\
&= \Phi\left(\frac{V_i - V_j}{\sigma} \right)
\end{aligned}
$$

where $\Phi(\cdot)$ denotes the standardized cumulative normal distribution. In the case of the logit model, the random parameter $\varepsilon \equiv \varepsilon_j - \varepsilon_i$ is logistically distributed and the probability for alternative *i* can then be solved as

$$P_i = \int_{\varepsilon=-\infty}^{V_i - V_j} f(\varepsilon) d\varepsilon$$

$$= \frac{1}{1 + e^{(V_i - V_j)}} \qquad (4.16)$$

$$= \frac{e^{V_i}}{e^{V_i} + e^{V_j}}$$

Since the probit model has no closed form, the coefficients are estimated through a series of simulations. Earlier this was seen as an analytical disadvantage, and hence the binary logit model, with its closed form, was often preferred due to its ease of manipulation. However, as modern computers have no problems in solving the simulated models, this argument is no longer valid. (Greene 2003) argues that there might be practical reasons for favouring one model over the other, but that it is difficult to justify the choice of one distribution over another on theoretical grounds. He further notes that the choice of model appears not to make much difference in most applications. (Ohler et al. 2000) are in agreement with this and add that the assumption of a normal distribution (probit) or a gumbel distribution (logit) is inconsequential, as the researcher needs thousands of tail observations to detect differences between the two distributions[21].

An issue that the researcher does need to be aware of, however, is that the probit and the logit models differ in the scaling of the coefficients. The logit model has a variance = $\text{var}(\varepsilon_j - \varepsilon_i) = \pi^2 / 6$, whereas the probit model has a variance = $\text{var}(\varepsilon_j - \varepsilon_i) = 1$. This implies that the logit coefficients are ≈1.3 times larger than the probit coefficients (the approximation works best at the centre of the distribution). It is important, therefore, to be aware of scaling when comparing the **absolute** estimates from the two models. This scaling issue is discussed again in section 4.4.6.

[21] What can matter though, is that the distribution of the random parameter in the logit model results in fatter tails than in the probit model (Ben-Akiva et al. 1997). Hence it can be argued that if the observed choices have 'extreme' values, e.g. that 95% of the respondents choose alternative *i* and 5% alternative *j*, the logit model will result in a more appropriate/realistic estimation, as the tails (and thus probabilities away from the middle value) are given larger weight.

4.4.2 Extension of binary models: Random effects

Respondents are most often asked a series of choice sets, which means that more than one observation is collected from each individual implying panel data observations. Some problems may arise using panel data, as 'normal' binary models cannot handle this kind of situation in which unobserved factors are correlated (Train 2003). Because respondents are asked to perform more than one choice, the within-individual variation across discrete choices may not be random. Therefore, a random parameter that is freely correlated within an individual but not across individuals is preferable. The random effects model incorporates such a parameter and hence increases data fit. For individual *n* this implies that utility is given by

$$U_{in} = V_{in} + \varepsilon_{in} + \mu_n \tag{4.17}$$

The error term ε_{in} is the random error term that includes random variation *across discrete choices*, while μ_n is the random error term *across respondents* and is constant for each individual. This means that ε_{in} is the error term due to difference among observations and μ_n is the error term due to differences among respondents - termed person-specific variation. Person-specific variation is present when there is some unobserved taste parameter that makes two otherwise identical individuals answer differently to the same choice. The μ_n thus captures the between-subject variability – also known as heterogeneity among individuals (Greene 2003).

The random effects model can be applied to both the logit model and the probit model. A review of the literature revealed that the random effects *probit* model is by far the most commonly applied binary model in DCEs. This can be explained by the fact that μ_n is assumed to be normally distributed and the random effects probit model thus seems intuitively correct as the two error terms then have identical distributions; this does not prevent the use of a logit specification, however. (Wooldridge 2002) points out that random effects logit is not as attractive as the random effects probit because there are no simple estimators available - integrating the logit response with respect to the normal density yields no simple functional form. For an example of the application of the random effects logit model in health economics, see (Hall et al. 2002).

4.4.3 Multinomial discrete choice models

In the case of multinomial choices, the derivation of useful choice models and appropriate estimation methods becomes considerably more complex than for binary discrete choice analysis. In particular, it is not sufficient simply to specify the univariate distribution of the differences in the disturbances. Instead, the complete joint distribution of all the random error terms have to be characterized (Ben-Akiva & Lerman 1985).

The MNL model, developed by McFadden (1974)[22], is by far the most used multinomial model and can with good reason be considered as the origin of multinomial models. The degree of estimation complexity increases rapidly as one moves away from MNL and relaxes the assumptions for the variance-covariance matrix. The MNL model has a special property as it assumes *independence of irrelevant alternatives* (IIA)[23]. This implies that the ratio of the probabilities of choosing one alternative over another is unaffected by the presence or absence of any additional alternatives in the choice set. The IIA assumption provides some clear advantages as it makes the MNL model very simple to operate. However, the IIA assumption also has some serious shortcomings. This is the case when observed and unobserved attributes of utility are not independent of one another and/or if the unobserved components of utility are correlated among alternatives, leading to biased utility parameters and forecast errors. Box 4.1 illustrates the problems related to the IIA property. As described above, the IIA restriction implies that the odds ratio of two alternatives i and j are the same, regardless of which other alternatives are available. It is straightforward to demonstrate that the IIA holds for the MNL model (Haan 2004):

$$\frac{P_i}{P_j} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}} \Bigg/ \frac{e^{V_{nj}}}{\sum_j e^{V_{nj}}} = \frac{e^{V_{ni}}}{e^{V_{nj}}} \tag{4.18}$$

Louviere (Bennett & Blamey 2001, chapter 2) argues that even with the strong assumption of IIA, the MNL model is still very useful and robust; the violation of IIA can be avoided by the inclusion of interaction variables such as sociodemographics. The assumption of IIA can be avoided by using a more complex model such as nested logit, multinomial probit and mixed logit. Due to the

[22] McFadden first called this model the conditional logit model as, in the multinomial case, it could be interpreted as the conditional distribution of demand given the feasible set of choice alternatives. Today, however, this model is more commonly called the MNL model (McFadden 2001).

[23] The IIA assumption is identical to the assumption of *independent and identically distributed* (iid) random components of each alternative

increased computer capacity and development of new models that solve the problem of IIA, these models (especially the mixed logit model) are gaining terrain, although the MNL remains a popular choice modelling framework. See Table 4.2 for an overview of the most prominent discrete choice models applied in the DCE literature.

Box 4.1. An illustration of the IIA property

Suppose that preferences for screening are to be examined. Besides different screening programmes (variation in attribute levels) it is obvious to include an "opt-out" option, i.e. that respondents may not wish to participate in the screening programme (having negative utility for the given programmes)[1]. For simplicity it is assumed that the representative utility for attending/not attending is the same, such that choice probabilities are equal: $P_A = P_N = ½$ (relative probabilities = 1:1); moreover the number of respondents is set at 60 individuals. Consider the case in which it is decided to include two screening programmes in the choice set, such that the respondents can choose between two screening programmes and no screening programme. Due to the assumed choice probabilities ($P_A = P_N = ½$) we will observe the following pattern of probabilities:

Screening 1	Screening 2	No screening
P=1/4	P=1/4	P=1/2
n=15	n=15	n=30

in which the relative probabilities: screening 1/screening 2/no screening = 1:1:2.

It is clear that screening 1 and 2 have some properties which make them more similar to each other than either is to the no screening option (i.e. they are correlated). This is what causes the problem: The MNL model does not take such pattern into account. The IIA assumption implies that the MNL model will treat all the alternatives the same and hence forces the following probability patterns on the observed choices:

Screening 1	Screening 2	No screening
P=1/3	P=1/3	P=1/3
n=20	n=20	n=20

Relative probabilities: screening 1/screening 2/no screening = 1:1:1

[1] The inclusion of the "opt-out" alternative is discussed in more detail in section 5.6.

Table 4.2: Overview of multinomial models

Discrete choice model	**Choice probability, P_{ni} =**	**Notes:**
Multinomial logit (MNL) – also termed conditional logit	$\frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}$ $j = 1,....,J$ $n = 1,.....N$	The most widely used and primary model. Its advantages include estimation simplicity, closed form specification and robustness. Assumes iid ➔ unobserved factors uncorrelated over alternatives and same variance for all alternatives, known as IIA. Violation of IIA causes bias. Some tests exist that make it possible to examine the accuracy of IIA (e.g. Train 2003 pp. 53). Does not account for correlation within each respondent's series of choices.
Multinomial probit (MNP)	$\int_{\varepsilon_n} \Phi(\varepsilon_n) d\varepsilon_n$	~N(0,Ω). The covariance matrix Ω accommodates for any pattern of correlation and heteroscedasticity and hence makes it possible to handle correlation over alternatives and time. This is the main advantage of using the probit. Functional limitation arises from the normal distribution assumption (density on both sides of zero).
Nested logit (NL)	$\frac{e^{V_{ni}/\lambda_k}(\sum_j (e^{V_{nj}/\lambda_k})^{\lambda_k - 1}}{\sum_{l=1}^{K}(\sum_j (e^{V_{nj}/\lambda_k})^{\lambda_l}}$	λ_k (named inclusive value/log-sum term)= measure of the degree of independence in unobserved utility among the alternatives in nest *k*. NL allows for correlation of error terms within one nest but not for any two alternatives in different nests (i.e. assumption of IIA within a nest)

Discrete choice model	Choice probability, P_{ni} =	Notes:
Heteroscedastic extreme value (HEV)[24]	$\int \left[\prod_{j \neq i} e^{-e^{-(V_{NI} - V_{nj} + \theta w)/\theta}} \right] e^{-e^{-w}} e^{-w} dw$	Allows the variance of unobserved factors to differ over alternatives, i.e. variance (the scale parameter) has a functional form.
Mixed logit[25] – Random parameters model	$\int \left(\frac{e^{V_{ni}(\beta)}}{\sum_j e^{V_{nj}(\beta)}} \right) f(\beta) d\beta$	Not subject to IIA, accommodates correlations among panel observations and accounts for uncontrolled heterogeneity in tastes across respondents. Allows the unobserved factors to follow any distribution. Unobserved factors decomposed into two parts: one part that follows any distribution and hence contains all the correlation and heteroscedasticity, and another part that is iid extreme value Type I distributed. In the statistic literature, the weighted average of several functions is called mixing distribution, hence mixed logit. Assuming that the mixing distribution is discrete in nature induces the *latent class models*[26].

Sources: (Ben-Akiva & Lerman 1985;Bennett & Blamey 2001;Louviere et al. 2000;Train 1993;Train 2003)

The nested logit model

The nested logit model is a generalisation of the multinomial logit model that allows for a particular pattern of correlation in unobserved utility (i.e. differences in cross-elasticities of substitution across alternatives)[27]. The nested logit model is thus appropriate when the set is of alternatives faced by a decision-maker can be portioned into subsets, called nests (Train 2003). A nested structure suggests that respondents initially choose between e.g. 'doing something' and 'doing nothing' (the 'upper' model/marginal probability) and then subsequently choose between the

[24] The HEV model and the nested logit model belong to the class generalised extreme value models
[25] The random parameter model can also be applied to binary choice tasks
[26] For an example of application of latent class model to DCE, see (Boxall & Adamowicz 2002)
[27] Full information maximum likelihood method is used to estimate model parameters in the nested logit model

different alternatives if the 'doing something' nest is chosen (the lower model/conditional probability); see Figure 4.5.[28]

The properties of the nested logit model are that within a nest the IIA holds, whereas IIA does not hold between nests. The nested logit model thus provides a way to link different but interdependent decisions, and to decompose a single decision to minimize the restriction of equal cross-alternative substitution. The nested logit model provides a way to identify the behavioural relationship between choices at each level of the nest, and also enables the researcher to test the consistency of the partitioned (nested) structure with random utility maximization (Louviere et al. 2000).

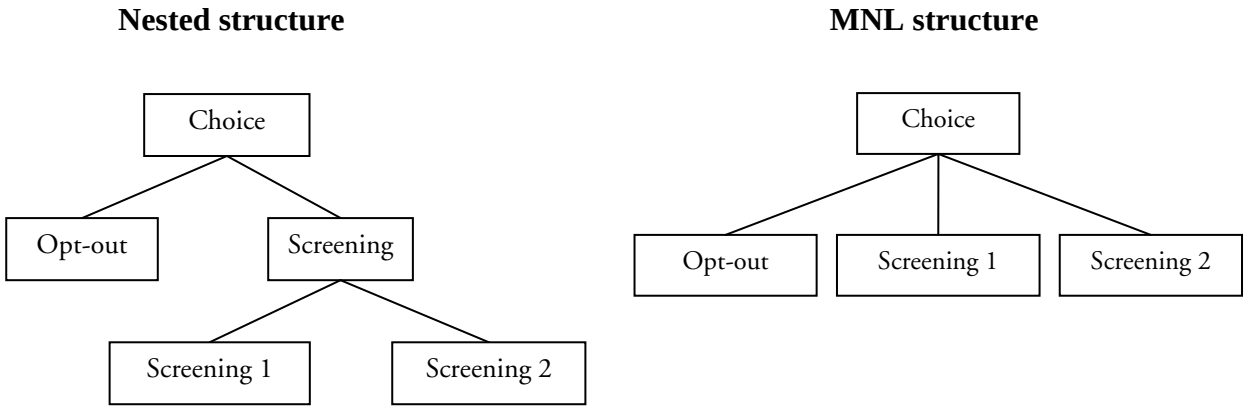

Figure 4.5: A comparison of the nested logit model structure and the multinomial structure

A special feature of the nested logit model is the following equation derived from the conditional probability:

$$I_{nk} = \ln \sum_{j \in B_k} e^{Y_{nj}/\lambda_k} \tag{4.19}$$

Rearranging with respect to λ_k

$$\lambda_k I_{nk} = \ln \sum_{j \in B_k} e^{Y_{nj}} \tag{4.20}$$

[28] For simplicity, the discussion of the nested logit model is restricted to the situation with two nests. The model can easily be extended

B_k denotes the subset of alternatives in the *k*'th nest. This index is commonly referred to as the *inclusive value* (IV) or, alternatively, the *logsum* or *expected maximum utility*, whereas λ_k is the coefficient of the inclusive value. The inclusive value defines a utility index associated with a partitioned set of alternatives and can be interpreted as the expected utility that respondent *n* recognizes from the choice among the alternatives in nest B_k. Moreover, it is λ_k that reflects the degree of independence among the unobserved portions of utility for alternatives in nest B_k. To be consistent with utility maximization, λ_k must lie between zero and one. When the coefficient tends towards one, the correlation among the unobserved components of utility for the alternatives within a nest decreases; and at the value of one, no correlation exists and the choice probabilities become an MNL. Testing the constraint λ_k=1 (by log-likelihood ratio testing) is thus the same as testing whether the standard MNL model is a better specification than the more general nested logit model. Conversely, if λ_k=0, no independence exists between the two nests and the decision-making can be said to be separated into two distinct strategies. This implies that no trading occurs between the two nests, and empirically this behaviour is revealed as each respondent consistently choosing either one of the two screening alternatives *or* the opt-out option. Moreover, (Train 2003) notes that, as $\lambda_k \rightarrow 0$, the nested logit approaches the 'elimination by aspects' model of Tversky (1972).

Ryan & Skåtun (2004) used a nested logit model in the examination of cervical screening. An opt-out option was included in their study such that respondents had the possibility of choosing between three alternatives (as illustrated in Figure 4.5.) The decision to undertake screening (or not) was found not to be dependent on the attributes of the alternatives, but rather on some sociodemographic variables (i.e. λ_k=0). This result indicates that the participation rate cannot be increased by improving the quality of the screening programme.

The random parameter logit

The random parameters model (RPL) is a specification of the mixed logit in which the coefficients are assumed to be random (in contrast to the 'error components' specification in which the interpretation of the model differs from the random parameters model). Considerable progress is being made in the application of more complex models such as the RPL model and there is greater availability of computer packages that include standard procedures for such models[29]. Although the RPL model is not yet well understood, it is likely that the number of DCE studies applying RPL

[29] To the author's knowledge, it is possible to run a random parameter logit in the following computer packages: Stata (e.g. Haan 2004), LIMDEP (e.g. Carlsson et al. 2003) and Gauss (e.g. Train 1998) with different limitations

will increase dramatically over the next years. The RPL appears to have some advantages over other discrete choice models as it provides the researcher with valuable information regarding the interpretation of the unobserved part of utility, and provides unbiased estimates even if unobserved heterogeneity is present in the data. Train (1998) noted that logit and nested logit specifications have many advantages, including simplicity of estimation, but that the same models also impose several and well-known restrictions that are not always desirable. Among these are:

- Coefficients of variables that enter the model are assumed to be the same for all respondents, i.e. that respondents have homogeneous preferences. This implies that respondents with the same observed characteristics (when accounting for these in the estimation) value the attributes equally.
- Logit and nested logit assumes IIA; this implies that the models necessarily predict that a change in the attribute of one alternative changes the probabilities of the other alternatives proportionally.
- Logit models (including nested logit) assume that unobserved factors are independent over time for each respondent. In a discrete choice experiment setting, this implies that unobserved factors are independent over the choice sets faced by each individual. However, it is likely that these unobserved factors that influence respondents choices are persistent (correlated) over choice sets. This assumption - along with the assumption of constant observed factors for each individual - gives rise to the assumption of stable preferences for each individual, i.e. the same tastes are used by the respondent to evaluate each choice set (alternative). This assumption is in line with the economic theory of rational behaviour[30].

A parameter that is found to be insignificant can be interpreted in two ways:

1. That the variable (the attribute) associated with the parameter did not influence the respondents' choices. This implies that this attribute (as it appears in the experiment with the ascribed levels) is not considered important to the respondents.
2. That preference heterogeneity exists. In this case, the attribute does affect respondents' choices, but with some respondents preferring one attribute level and some respondents

[30] This individual effect is the same as the effect accounted for in a random effects specification for binary choices.

preferring a different attribute level (and so forth depending on the number of levels). The attribute level effects off-set each other and result in an insignificant parameter estimate.

The advantage of the random parameters model is its ability to separate these effects and thus to allow correct interpretation of insignificant parameters (e.g. Train 1998).

Working through the RPL step by step (based upon Hensher & Greene 2002;Train 2003), we know from equation 4.1 that utility can be divided into two parts: an observable and an unobservable part,

(4.21) $$U_{ni} = V_{ni} + \varepsilon_{ni} = \beta x_{ni} + \varepsilon_{ni}$$

Instead of assuming that β is fixed, β is assumed to vary among respondents, hence letting $\beta_n = \beta + \theta_n$ where β is the mean of the coefficient and θ_n a random term that captures non-observable individual effects, such as tastes. The utility function now becomes

(4.22) $$U_{ni} = V_{ni} + \varepsilon_{ni} = \beta_n x_{ni} + \varepsilon_{ni} = (\beta + \theta_n) x_{ni} + \varepsilon_{ni}$$

Assuming that ε_{ni} is iid extreme value type 1, the probability for choosing alternative *i* thus becomes

(4.23) $$L_{ni} = \frac{e^{V_{ni}(\beta)}}{\sum_j e^{V_{nj}(\beta)}} = \frac{e^{\beta_n x_{ni}}}{\sum_j e^{\beta_n x_{nj}}}$$

If θ_n was zero, then β_n would be fully known and the random parameters model would collapse to the standard logit model. If, however, the researcher does not know the respondents' individual tastes, then the coefficients vary in the population with density denoted $f(\beta_n \mid \beta, \theta)$ where β denotes the mean and where θ are the parameters of this distribution (often representing the standard deviation of tastes in the population). Since the researcher does not observe the actual tastes, the probability becomes the integral of L_{ni} over all possible values of β, weighted by the density of β. The unconditional choice probability is therefore the integral of the logit specification over all the possible values of β_n:

$$(4.24) \qquad P_{ni} = \int L_{ni} f(\beta) d\beta = \int \left(\frac{e^{\beta_n x_{ni}}}{\sum_j e^{\beta_n x_{nj}}} \right) f(\beta) d\beta$$

The probability is approximated through simulation for any given value of θ, i.e. for a given value of the parameters θ, a value of β is drawn for its distribution (named β^r, *r* indicating the *r*'th draw). This procedure is repeated for many draws, and concluded by averaging the result. The average is then the simulated probability (*SP*)

$$(4.25) \qquad SP_{ni} = \frac{1}{R} \sum_r L_{ni}(\beta^r)$$

where *R* is the number of random draws[31]. The simulated probabilities are inserted into the log-likelihood function. In the case where each respondent is asked more than one choice – which is the case in many DCEs - it is standard to treat the coefficients as varying over respondents but being constant over choice situations (the number of choice situations denoted *t*) for each respondent. This is the same as assuming that respondents have stable preferences and implies that the utility function and unconditional probability for a sequence of choices becomes (Revelt & Train 1998)

$$(4.26) \qquad U_{nit} = \beta_n x_{nit} + \varepsilon_{nit}$$

$$(4.27) \qquad P_{nit} = \int L_{nit} f(\beta) d\beta = \int \left(\frac{e^{\beta_n x_{nit}}}{\sum_j e^{\beta_n x_{njt}}} \right) f(\beta) d\beta$$

A key element in the use of the RPL model is the assumption regarding the distribution of each of the random coefficients. This is not straightforward, but involves great consideration as to

[31] Train (2003) discusses the possibilities for applying sequences of draws that are not purely random; these include antitethics, systematic sampling and Halton sequences. He argues that these drawing sequences can provide better approximations to the integral than a sequence of purely random draws.

the most intuitive approach. Furthermore, as noted by Carlsson et al. (2003), the choice is often limited by difficulties of model estimation and the availability of econometric software. The most common functional forms are Hensher & Greene (2002):

- Normal
- Log-normal
- Triangular
- Uniform

Normal and log-normal distributions are particularly popular. The log-normal distribution is especially appropriate for attributes that tend to take a specific sign – such as the cost attribute – as this distribution only lies above zero (the use of a normal distribution for a price coefficient would imply that some proportion of the population prefers higher prices). Uniform distribution is a sensible approach for dummy variables[32], whereas triangular (flat on the top) and normal distributions are appropriate for attributes where it is expected that respondents will differ in having positive and negative preferences for the attributes (with the triangular distribution having a higher percentage that is negative).

In practice, researchers often find that any distribution has strengths and weaknesses. For instance, a special problem related to the application of the log-normal distribution is that the tail is unbound (compared to e.g. the normal distribution). Applying such a distribution to the cost attribute would result in some individuals being assumed to have unlimited WTP. This aspect is not desirable as it is behaviourally unrealistic and conflicts with the idea of compensatory decision-making. A possible solution to this problem (and to restrictions related to the other distributions) is the imposing of constraints on the distributions, i.e. making bounds. Bounding the normal distribution, for example, in such a way that it can only take values of one sign makes it suitable for attributes in which preferences are only positive or negative; the bounded normal distribution in this case has an advantage over the log-normal distribution, but is at the same time bounded. The procedure for distributional restrictions entails numerical difficulties which until now have limited the application of such methods (Train & Sonnier 2003).

[32] With a bound (0,1) (Hensher & Greene 2002)

McFadden & Train (2000) showed that any random utility maximization model[33] can be approximated by the mixed logit model with an appropriate choice of variables and mixing distribution. This characteristic of the mixed logit might be taken as a reason for always choosing such a model – for instance in comparison to the nested logit. Train (2003) warns that such model replacement should be made with caution as it reduces accuracy. In the case of using a mixed logit to represent the substitution patterns of a nested logit, the closed-form integral of the nested logit is replaced with an integral that needs to be simulated. Train suggests that the only advantage of the mixed logit is in situations where numerous nesting structures are to be tested and where some coefficients are assumed to be random. McFadden & Train (2000) have developed a test that specifies the need for mixing. The test is made solely on MNL model estimates. Two MNL models are compared: Model 1 containing the attribute variable and Model 2 containing the attribute variables and artificial variables, z_i, for selected components of x_i

$$z_i = \frac{1}{2}(x_i - \sum_j x_j P_j) \tag{4.28}$$

A Wald chi or Likelihood Ratio test is then used for the hypothesis that the artificial variables should be omitted from the MNL model. If the hypothesis is rejected, mixing is needed.

In contrast to environmental economics, the use of more complex models, such as the RPL model, has been limited in health economics (Hanley et al. 2003). The random effects probit model is by far the most used model in health economics. It allows for multiple observations for an individual but is restricted in the sense that it only allows binary responses. Johnson et al. (2000) use an RPL model in their estimation of willingness-to-pay for improved health for respiratory and cardiovascular patients. (For a very readable DCE article using an RPL model (in environmental management), see Carlsson et al. (2003)).

Hensher & Greene (2002) provide a comprehensive discussion of the RPL model, including some worked examples. They argue that there are at least ten key issues to consider in specifying, estimating and applying an RPL model (most of these have been mentioned in this paper; for extended reading, see Hensher & Greene (2002). These issues are:

[33] Discrete choice models in line with random utility theory

1. **Selecting the parameters that are to be random**
2. **Selecting the distribution of the random parameters**
3. **Specifying the way in which random parameters enter the model** (for example for attributes that possess disutility and which are assumed to be log-normally distributed (e.g. the cost attribute), the sign needs to be reversed prior to model estimation)
4. **Selecting the number of points on the distribution** - relates to the number of draws and how the draws are obtained
5. **Decomposing mean parameters to reflect covariate heterogeneity** (including interaction terms in the estimation)
6. **Empirical distributions (**investigating the distribution empirically prior to assuming an distribution)
7. **Accounting for observations drawn from the same individual** - panel data implications; correlation of choice situations associated with each individual
8. **Accounting for correlation between attributes** - correlated alternatives and choice situations usually go hand in hand
9. **Taking advantage of priors in estimation and posteriors in application** - the use of Baysian methods
10. **Willingness-to-pay challenges** - random parameters increase the complexity of WTP estimates as they possible require incorporation of assumed distributions. When deriving WTP estimates, the researcher can use all the information in the distribution or just the mean and standard deviation. (This issue is discussed further in section 4.8).

Although the RPL model possesses some real advantages compared to other models, it is also important to remember its shortcomings. As noted by Hensher and Greene:

> *"[….] despite great progress in developing ever more powerful and complex models that can capture many aspects of choice behaviour, it nonetheless is the case that such models are only as good as the data from which they are estimated"* (Hensher & Greene 2002, pp 24)

Louviere et al. (2002) point towards an extension of the RPL model in which not only the parameter but also the random error term associated with each choice option has a unique systematic and random component for each individual. They argue that it makes more sense behaviourally to assume that the systematic as well as the random component is distributed.

4.4.4 Alternative specific constants

The constants in discrete choice modelling are named alternative specific constants (ASC). If the choice set constitutes *n* alternatives, then one needs to specify *n-1* alternative specific constants, with one of the constants normalized to zero. It is irrelevant which of the constants is normalized to zero as the other constants are interpreted as being relative to the constant set to zero (i.e. the one alternative not represented by a constant is the comparator). The ASC for an alternative captures the average effect on utility of all factors that are not included in the model. Thus they serve a similar function to the constant in a regression model (OLS), which also captures the average effect of all non-included factors (Train 2003). Moreover, the ASC has the ability to capture non-participation when applied to nested logit models (Hanley et al. 1998).

Viewed from a theoretical perspective, however, it is not always appropriate to include an ASC. Thus raises the question: When is it appropriate to include a constant(s)? And when is it not appropriate? If the alternatives are *generic* (implying that they only differ on the attributes and are not labelled in any way), then the ASC is assumed to be zero as the difference in utility between the alternatives is caused only by the attributes; this is already incorporated in the model (remembering that only differences in utility matter). If the researcher in such a case chooses to include a constant, then this constant only symbolizes the tendency, for instance, to choose the right alternative, all other things being equal; and we would expect the constant to come out insignificant. On the other hand, if the alternatives are labelled in some way that might influence differences in utilities between the alternatives, then ASC have to be included. In this case, it is important that the researcher remembers to effects code the qualitative attributes. To summarize:

Situations in which it is appropriate to include ASC:

- Generic alternatives. When alternatives only differ in attribute levels and thus have no special features attached to them

Situations in which it is appropriate to include ASC:

- Inclusion of a constant comparator (alternative): One alternative is the same for all choice sets. This is the case for the inclusion of opt-out/status-quo alternatives
- Labelling of alternatives: When alternatives possess utility beyond that ascribed by the attributes, e.g. brand names

4.4.5 Estimation procedure – the likelihood function

The estimation of the choice models is most often based on the method of maximizing the likelihood function[34]. The maximum likelihood estimation procedure is very burdensome and is virtually impossible to apply without a computer. The method is based on the idea that a given sample could be generated by different populations, and is more likely to come from one population than another. Thus, the maximum likelihood estimates are that set of population parameters that generates the observed sample most often (Louviere et al. 2000). Consider the likelihood of any sample of observations. Since they, by assumption, are drawn at random from the whole population, the likelihood of the entire sample is the product of the likelihood of the individual observations (Ben-Akiva & Lerman 1985). The coefficients are estimated by the maximization of likelihood function.

Numerous computer packages contain routines for the estimation of discrete choice models with linear-in-parameters representative utility (Train 2003). Among these are SAS, Stata, Gauss and Limdep.

The likelihood ratio test

As with regressions, standard t-statistics are used to test hypotheses about individual parameters in discrete choice models. For more complex hypotheses, however, a likelihood ratio test (LL-test) can be used. The LL-test is used in the same way that an F-test is used in ordinary least square estimation. The test can for instance be used to test for the null hypothesis that all the parameters are zero; the test is not so useful for this purpose, however, because the null hypothesis is almost always rejected at a low level of significance. The most useful application of the LL-test is

[34] Other functions that can be used in the maximization procedure include the simulated likelihood function or squared moment conditions. Furthermore, there are estimation procedures that do not require maximization of any function; the Baysian procedures being the most prominent.

for more specific hypotheses such as testing the null hypothesis that several parameters are zero, that two or more parameters are equal or as a model selection criteria (using the same data).

The test statistic is

$$\text{LL} - \text{test} = -2 * (\text{LL}(\beta^R) - \text{LL}(\beta^U)) \tag{4.29}$$

Where β^R denotes the restricted maximum value of the likelihood function under the null hypothesis and β^U denotes the unrestricted maximum of the likelihood function. The statistic used is chi-squared distributed with (K_U-K_R) degrees of freedom, where K is the number of estimated parameters. If the value of the LL-test exceeds the critical chi-squared value then the null hypothesis is rejected. Note that it is only possible to compare log-likelihood estimates for models that share common distributional assumption, e.g. the MNL model versus the more general RPL model.

Goodness of fit

In contrast to the linear regression model, there is no single measure for the goodness of fit in discrete choice models. One of the most well-known measures of goodness of fit is the McFadden R^2 (also termed pseudo R^2). It is defined as follows:

$$R^2 = 1 - \frac{\text{LL}(\beta^R)}{\text{LL}(\beta^U)} \tag{4.30}$$

The nominator is the value of the log likelihood function at the estimated parameters and the denominator is its value when all the parameters are set equal to zero. If the estimated parameters are irrelevant then the nominator equals the denominator and so $R^2 = 0$. Similarly, $R^2 = 1$ when the respondent's choice can be perfectly predicted. The larger the difference between the two log likelihood values, the more the extended model adds to the very restrictive model (Verbeek 2000). It is important to note that the McFadden R^2 (and other goodness of fit measures for discrete choice models) is not comparable with the OLS R^2 in its interpretation. The OLS R^2 explains the degree to which the dependent variable is explained by the estimated model. The McFadden R^2 has no intuitive meaning for values lying between the two extremes ($\rho = 0, 1$). It is the percentage increase

in the log likelihood function above the value taken at zero parameters. However, the meaning of such a percentage change is not quite certain. A valid and possible application of the measure is in the comparison of several models with the same data and with the same set of alternatives. In this case, the model with the highest R^2 fits the data best and hence is to be preferred to the others (Train 1993). In some circumstances it is appropriate to apply adjusted goodness of fit measures that penalizes the loss of degrees of freedom that occurs when a model is expanded. Akaike information criterion (AIC) and Baysian information criterion (BIC) are two such measures (Greene 2003).

Another goodness of fit statistic is the *percent correctly predicted*. The alternative with the highest probability is identified for each respondent and it is thereafter determined whether or not this was the alternative chosen. Train (1993) and Verbeek (2000) argue, however, that this statistic provides even less information about the model than the McFadden R^2 and, in addition, that it misinterprets the nature of probability[35].

4.4.6 Combining of data - and scaling

When applying DCEs, it may sometimes be appropriate to combine the data in some way in order to obtain additional information. Combinations of data include

- Different sources of SP data (e.g. Cameron et al. 2002)
- Segments/sub-sets within a single data set (e.g. Saelensminde 2001)
- Stated preference data (SP) & revealed preference (RP) data – Choice data are especially appropriate for data combining as they in nature possess variances in data (due to use of attributes and levels)

There are several reasons for merging data, of which the most important are (Bateman et al. 2002):

- A verification of convergent validity
- As a means of more efficient sampling
- To combine the desirable features of two approaches
- To examine methodological and theoretical issues

[35] Imagine a model that predicts a two-alternative situation with probabilities 0.9 and 0.1, respectively. If the situation was repeated 100 times, the researcher's best prediction of alternatives chosen would be 90 and 10. The statistic is based on the notion that the best prediction of alternatives in each situation is the alternative with the highest probability. This implies that the alternative with probability 0.9 should be chosen all 100 times.

The combining of data is not as straightforward as might first be expected. This is foremost due to the fact that, in the estimation of parameter vectors, these vectors are confounded with the variance of the unobserved effects. For the same reason, parameter vectors from different models and/or data sets are not comparable in absolute terms. As mentioned earlier, probit and logit estimates cannot simply be compared as the models by distributional nature differ in scale by a factor $= \pi / \sqrt{6}$.

Besides this scale factor, there are other reasons why variance might differ - even applying the same model. Studies have shown, for example, that variance tends to be smaller for RP data than for SP data. Train (2003) mentions an example in which variance tends to differ between two groups of respondents (from different cities) due to heterogeneity. In both cases, parameter estimates – and thus the underlying preferences – might turn out to be identical even though they at first glance look very different. This is due to differences in scale. The bottom line is that comparisons of model parameters from different sources of data must take into account potential differences in the variances of the error components - otherwise it is not possible to separate variance from parameter estimates.

For the MNL model, the relations between variance and scale and parameters are

(4.31) $$\sigma^2 = \frac{\pi^2}{6\mu^2} \quad \Rightarrow \quad \sigma = \frac{\pi}{\sqrt{6}\mu}$$

and

(4.32) $$\beta = \frac{\beta^*}{\sigma} = \beta^* (\frac{\pi}{\sqrt{6}})\mu$$

hence giving the choice probabilities

(4.33) $$P_{in} = \frac{e^{\mu\beta x_{ni}}}{\sum_j e^{\mu\beta x_{nj}}}$$

This implies an inverse relationship between the scale factor and the random component variance. The parameters are estimated, but for interpretation it is useful to recognize that these estimated

parameters are actually estimates of the 'original' coefficients multiplied by the scale parameter; see Figure 4.6 for the effect of the scale parameter on choice probabilities. The coefficients that are estimated indicate the effect of each observed variable *relative to* the variance of the unobserved factors. A larger variance in unobserved factors leads to smaller coefficients, even if the observed factors have the same effect on utility (Train 2003). As the scale parameter is constant in an MNL (and other models) it does not affect the utility orders in separate data sources[36]. However, the scale parameters may differ significantly between data sets. As the scale parameter is confounded with the parameter vector, the absolute scale parameter cannot be identified in a single data source. Rather, ratios of μ's can be estimated relative to an arbitrary reference source (Severin et al. 2001). An example of a relative scale parameter between two sources is illustrated in the equation below (for the MNL model):

$$(4.34) \qquad \frac{\sigma^2_{data1}}{\sigma^2_{data2}} = \frac{\pi^2/6\mu^2_{data1}}{\pi^2/6\mu^2_{data2}} = \frac{\mu^2_{data2}}{\mu^2_{data1}} = \left(\frac{\mu_{data2}}{\mu_{data1}}\right)^2 \approx \left(\frac{\hat{\mu}_{data2}}{\hat{\mu}_{data1}}\right) = \left(\frac{0.9}{1}\right)^2 = 81\%$$

This is the same as saying that the variance of dataset 1 is about 81% of that of the variance of dataset 2, and hence is smaller. Different scale parameters can thus imply different data source variability. It is important to note that the scale parameter does not affect the ratio of any two coefficients, since it drops out of the ratio. Only the interpretation of the magnitudes of all coefficients is affected.

[36] As with the MNL model, similar identification (scale versus parameters) exists for other choice models such as the nested logit, binary models and multinomial probit model (Hensher et al. 1999).

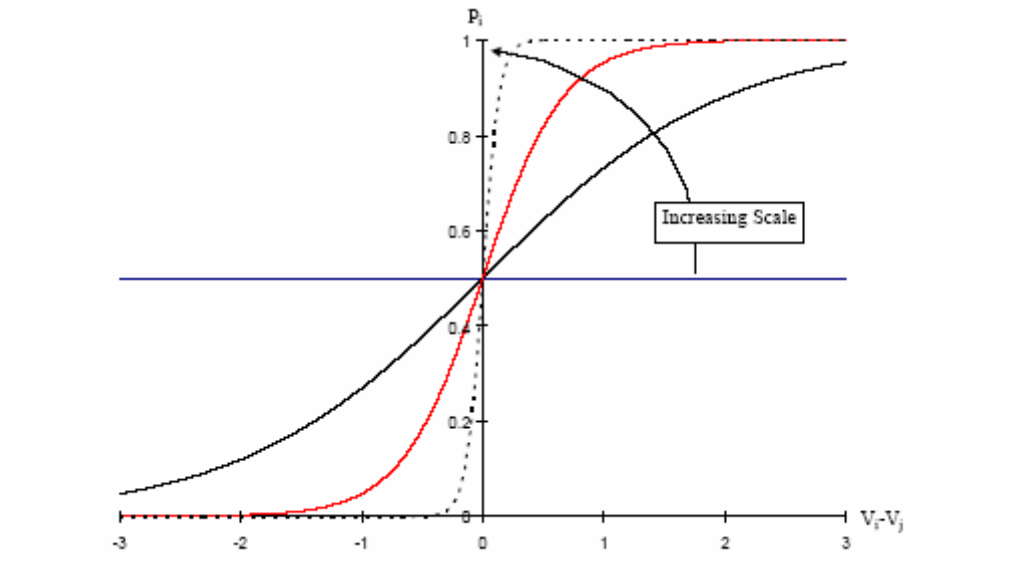

Figure 4.6: The effect of the scale parameter on choice probabilities. Source: (Adamowicz et al. 1998)

Suppose that we want to combine two data sets. The scale of one data set is normalized to 1 so that μ reflects the variance of the unobserved factors in data set 2 relative to data set 1. This implies that the utility for data set 1 and 2, respectively, is

Data 1: $U_{nj} = \beta x_{nj} + \beta_j^1 + \varepsilon_{jn}$

Data 2: $U_{nj} = \frac{\beta}{\mu} x_{nj} + \frac{\beta_j^2}{\mu} + \varepsilon_{jn}$

β_j denotes the alternative specific constants and, depending on the data sources, these might differ for the two data sets. If there is little variance in data 1 (which contains, for example, RP data), then the parameters will be determined primarily by data 2 (which contains, for example, DCE data). To the degree that data 1 contains usable variation, this information will be incorporated into the estimates (Train 2003). If response means are constant across two data sources but the variance of the random components differs, RUT predicts that the estimates of response mean effects will be proportional to one another, and the proportionality constant (the scale factor) will be the ratio of the variance in data source 1 relative to the variance in data source 2 (Louviere 2001b).

The potential problem of differences in scale factors between data sources was discovered by researchers at the beginning of the 1990s (Ben-Akiva & Morikawa 1990;Hensher et al. 1999;Swait & Louviere 1993). Severin et al. (2001) noted that this discovery might mean that earlier reported differences in choice model parameters may be more consistent with random component variance differences between segments than with real segment parameter differences. Severin et al. (2001) are in line with Louviere:

> "*Specifically, parameters recovered by all choice and ordinal response models really estimate* β/σ *[….] For example, since attraction effect researchers have not separated response means from response variability, claims that regularity is violated are suspect because target manipulation can affect response means, response variability or both. That is, if target manipulations impact response variability, this may shift distributions in ways that appear to violate regularity [….] Thus failures of rationality or economic theory that many consider well established may turn out to be not nearly so*" (Louviere 2001b, pp 507)

Two methods exist for finding the relative scale parameter(s): a manual process and as a simultaneous process of the regression analysis. Train (2003) notes that the latter is not feasible in most standard estimation packages, but points out that the researcher can easily modify available codes (or his/her own code) to allow for this extra parameter. Louviere et al. (2000) describe a simultaneous method using the nested logit model (in which the inclusive value relates to the scale parameter). The simplest method (proposed by Swait & Louviere (1993)) estimates the desired model parameters and the relative scale factor by manual search. The procedure yields consistent but inefficient estimates of the scale and the parameters as the standard errors might be underestimated, resulting in inflated *t*-statistics; on the other hand, the method has the undeniable and worthy virtue of simplicity. No matter which method is used for estimating the relative scale parameter(s), the underlying procedures are the same. The data is merged (pooled) under the assumption of equal parameter vectors, which implies that preferences are assumed to be identical across samples, and one can determine the scale parameter(s) which optimizes the log-likelihood function of the model fitted to the pooled data sets. The scale parameter finds the best factor for which the parameters of the two data sources are equalized. When variance differences are taken into consideration, the assumption of parameter equality is then tested; thereafter the hypothesis of equal scale parameters is tested. Following Swait & Louviere (1993), the tests to be performed are:

H_{1A}: $\beta_1=\beta_2=\beta$

H_{1B}: $\mu_1 = \mu_2$

in which 1 and 2 denote the individual data sets[37]. The test to be used is a log-likelihood ratio test, in which the statistics follow an asymptotic chi-square distribution:

H_{1A}: $\lambda_A = -2\left[LL_\mu - (LL_1 + LL_2)\right]$ with (the number of parameters +1) degrees of freedom

H_{1B}: $\lambda_B = -2\left[LL_P - LL_\mu\right]$ with (the number of scale factors +1) degrees of freedom

If H_{1A} is rejected, pooling of the data is not allowed as it cannot be determined whether either the parameter vectors or the scale *or* both are impacted. If H_{1A} is accepted, it is possible to pool the data, taking the relative scale parameter into consideration, however, if H_{1B} is rejected.

Combining revealed and stated preference data

Over the last decade the combining of SP data and RP data (also termed data enrichment) has gained increasing attention – especially in transportation and environmental economics (e.g. Boxall et al. 2003;Haener et al. 2001). The use of joint models in health care has been very limited. In a recent publication, Mark & Swait (2004) report the application of a joint model (nested logit model) to study physicians' prescribing patterns in the treatment of alcoholism. The physicians were asked questions related to the actual use of treatments and medication, while in a choice task they were asked to choose between two types of treatment (or no treatment), where the treatments differed in respect to efficiency, side effects, mode of action, route of administration and price. The authors reported that the stated preference data greatly improved the information attainable. This is due to DCE allowing estimation of attributes that are not in the market, that do not vary in the market or that are collinear with other attributes in the market. The results suggested that the reason for the physicians' infrequent use of medication was the poor quality of the products available.

The combining of data appears to possess some advantages as the SP data provide the needed variation in attributes, whilst the RP data ground the predicted shares in reality; the desirable features of the two approaches are thus combined. Other reasons for combining RP and SP data

[37] Here the test is shown for the combining of two data sets (which is the usual case); it can easily be extended to include more datasets, however.

include examination of convergent validity and increased sampling efficiency (Bateman et al. 2002). Haener et al. (2001) point out that within-sample predictive ability of joint models exceeds that of models estimated with RP data only, and that these models might be particularly appropriate for benefit transfer.

To utilize the relative strengths of the joint model, an estimation procedure is needed that allows the ratios of coefficients to be estimated primarily from the stated preference data, while allowing the alternative-specific constants and overall scale of the parameters to be determined by the RP data. This distinction allows the researcher to avoid many of the biases that SP data might exhibit. For example, respondents often say that they will buy a product far more than they actually end up doing so. The average probability of buying the product is captured in the alternative-specific constant for the product. If this bias exists, then the estimated constant for the SP data will be greater than that for the RP data. In a forecasting context, the researcher can use the constant from the RP data, thereby grounding the forecast in a market-based reality. Similarly, the scale for the RP data (which is most often the one that is normalized to 1) can be used in forecasting instead of the scale from the SP data, thereby correctly incorporating the real-world variance in unobserved factors (Train 2003).

Procedures for estimating discrete choice models on a combination of stated preference and revealed preference data are described by Ben-Akiva & Morikawa (1990), Hensher et al. (1999) in the context of logit models, and by Bhat & Castelar (2002) for mixed logit.

Studying the variance of the random component

Information on the variance of the random component can make a valuable contribution to understanding respondents' behaviour. Depending on which other factors have been taken into consideration, the scale parameter can provide information on heterogeneity and decision-making behaviour, such as departures from the assumption of rationality due to factors such as task complexity and context dependency. Saelensminde (2001) used the scaling procedure to investigate inconsistency in choices (i.e. testing the hypothesis that variance increases by the degree of inconsistency) and fatigue/learning effects (i.e. testing the hypothesis that the problem with multiple responses is not correlation, but non-stationary variance). He found that the incorporation of scaling with respect to the degree of inconsistency improved the model fit. Dellaert et al. (1999) examine the relationship between consistency (measured as variance differences) and differences in price level ranges. Findings indicate that choice difficulty increases as price range increases, explained by

the fact that trade-offs between the defined choice alternatives and the 'opt-out' alternative become harder to make because utility benefits of the good under consideration less clearly outweigh their costs.

Louviere (2001b) discusses the nature of the random component and its variance and argues that the quality of behavioural inferences inherently depends on the variance of the random component – which he calls *response variability* (also termed *unobserved variability*). This implies that both response variability and response means (parameters) drive response probabilities. The random component comprises components of response variability such as within-subjects, between-subjects, between-contexts, between-measurement instruments and between-time periods variability. In general, there is increased focus on the variance of the random component of utility. Louviere et al. (2002) discuss the importance of the response variability in random utility modelling and suggest five reasons for researchers to be interested in and concerned about response variability. These are: development of more general models, improving parameter estimates, forecasting and simulation, testing of the theoretical foundation, and combining of data sources. Like Louviere, they argue that unobserved variability does not only comprise unobserved heterogeneity, but also many other sources of variability such as condition, context and time dependency. What is valuable about RUT is that it provides a unified theoretical framework to formulate and test hypotheses about differences in response variability and response means on a level playing field, such that behavioural outcomes can be analysed as a function of manipulated variables, response modes, survey experiments, places, time periods and the like (Louviere 2001b). The examination of the random component is closely related to the earlier discussion of the attempt to include knowledge about psychological factors that affect decision-making into the valuation techniques of DCE. Louviere notes that optimally efficient statistical DCE designs cannot be developed without taking response variability into account a priori.

4.4.7 Measuring welfare

One of the strengths of DCE is its ability to evaluate various alternatives from one application, due to the differences in attribute levels. With the inclusion of a cost attribute in the choice set, it becomes possible to evaluate improvements and the supply of goods by means of welfare measurement. Welfare measures are especially appropriate for policy purposes, as they enable the researcher to perform analyses that estimate changes in welfare associated with a particular policy.

There are two main categories of welfare measures, which arise from two different approaches to using DCE in applied studies. These are (Bennett & Blamey 2001):

- 'State of the world' approach
- 'Multiple alternatives' approach

The 'State of the world' approach

The 'State of the world' approach refers to the estimation of welfare measures in which a given alternative will be consumed with **certainty**. This approach is thus in line with traditional neoclassical welfare theory, in which situations occur with certainty. This can be the case in the following situations:

- The current good (status quo alternative) will be replaced by another good (alternative)
- A new good is introduced to a market where no other varieties (alternatives) of the good currently exist
- For public goods

If, for example, a government decides to launch a new programme for air pollution protection and wants to evaluate the characteristics of such a programme (but not whether or not it should be promoted), then a 'state of the world' welfare measure would be the appropriate approach – because individuals cannot choose not to consume the good (air). A second example is the evaluation of an improved cancer treatment, such as chemotherapy. The new and better treatment will replace the old treatment and will become the only treatment available for this group of patients. Assuming that all patients attend for treatment, a 'state of the world' welfare measure would again be the appropriate approach.

Two types of welfare measures exist: welfare measures derived from the Hicksian demand curve (compensating and equivalent variation) and consumer surplus derived from the Marshallian demand curve (see Appendix I for an illustration). Using compensating variation (CV) as a basis, the definition is: The amount of money that has to be taken from (or given to) an individual in the new state (1) in order to keep him at the initial state utility level (i.e. indifferent between the new and the initial state). This implies

(4.35) $$U^0(m, p^0, x_j^0) = U^1(m - \mathrm{CV}, p^1, x_j^1)$$

where V^0 denotes the initial utility level and V^1 denotes the utility after a change from level of quality x_j^0 to x_j^1

Assuming that the price function is linear and that the marginal utility of money is constant, implies that:

(4.36) $$U = \gamma(m - p_j) + \beta x_j$$

and thus

(4.37) $$U^0 = \gamma(m - p^0) + \beta x_j^0 = \gamma(m - p^1 - \mathrm{CV}) + \beta x_j^1 = U^1$$

where γ is the price coefficient denoting marginal utility of income (or reverse marginal disutility of price). Assuming linearity in price implies that income cancels out and hence there is no income effect. As income cancels out, the price coefficient implicitly denotes marginal disutility of price (hence the negative sign)

(4.38) $$U^0 = -\gamma p^0 + \beta x_j^0 = -\gamma(p^1 + \mathrm{CV}) + \beta x_j^1 = U^1$$

In this case the estimation for a change in welfare becomes

(4.39) $$\Delta\mathrm{CV} = \frac{(\partial U / \partial x_j)\Delta x_j}{\partial U / \partial(p)} = \frac{(U^1 - U^0)}{\gamma} = \frac{\Delta\beta x_j}{\gamma}$$

In DCE it is standard to estimate a linear function of the attribute variables, including a cost attribute such as price for the particular good. For alternative *j*

(4.40) $$V_j = \beta_{price}\, p + \beta x_j$$

The researcher is normally interested in any changes in welfare due to a change in the qualitative attributes (i.e. holding price constant). WTP(CV) becomes the difference in utility between the two alternatives in question divided by the negative of the price coefficient (to change the term into marginal utility of income rather than marginal disutility of price (Louviere et al. 2000)

(4.41) $$\partial V = 0 \quad \Rightarrow \quad \Delta \mathrm{CV} = \frac{(\partial V / \partial x_j)\Delta x_j}{-\partial V / \partial(p)} = \frac{(V^1 - V^0)}{-\beta_{price}} = \frac{\Delta \beta x_j}{-\beta_{price}}$$

The welfare measure indicates the average WTP for a quality improvement in the good in a 'state of the world'. If it is a fact that the individual will definitely choose the good in question then any changes to the good can be assessed using the above formula – under the given assumptions. If instead the functional form of the price was, for example, characterized as a log-normal

(4.42) $$V_j = \gamma \ln(m - p) + \beta x_j$$

then marginal WTP for a change in one of the attributes and the welfare measure for a change in the alternative become

(4.43) $$\mathrm{MWTP} = \frac{\partial V / \partial x_j}{-\partial V / \partial p} = \frac{\beta(m - p)}{-\gamma}$$

(4.44) $$\Delta \mathrm{CV} = \frac{(\partial V / \partial x_j)\Delta x_j}{\partial V / \partial(m - p)} = \frac{(V^1 - V^0)}{u_{income}} = (m - p) - (m - p) * \exp\left(\frac{\beta \Delta x_j}{\gamma}\right)$$

Note that in this case income does not cancel out and hence there is an income effect.

While equation (4.39) provides a measure of economic welfare in cases in which the alternative is chosen with certainty (i.e. where the random component of utility in the RUT framework can be ignored), it does not address situations that involve many alternatives. In such situations, 'multiple alternatives' welfare estimation is needed.

The 'multiple alternatives' approach

The 'multiple alternatives' approach means that welfare measures are estimated on the grounds of **uncertainty**, i.e. the new alternative will be consumed with uncertainty. Only a fraction of the consumers will choose the improved/new alternative – there being consumers who still choose the alternative; previous non-demanders; and consumers who beforehand chose a substitute. Although the choice rate might increase for the given good, it will not achieve 100%. This is the case when alternatives are not mutually exclusive, i.e. when the consumer has the opportunity to 'shop' between different alternatives (substitutes exist) or to choose none of the alternatives. All private goods belong to this category and hence 'multiple alternatives' is the appropriate procedure in most cases.

The 'multiple alternatives' approach was developed by Small & Rosen (1981) (for application of discrete choice modelling in general) and Hanemann (1984) (for application of dichotomous choice CVM). In a specific choice setting, the researcher only knows that respondents have a certain probability of choosing a particular alternative, and hence welfare measurement needs to be extended to cases in which individuals have probabilities of choosing alternatives. The solution is to weight each alternative by the probability that it will be selected. This is analogous to the concept of expected utility in which the utility of being in several states of the world is weighted by the probability that each occurs, and the weighted utility sum equals the expected utility (Louviere et al. 2000, chapter 12).

The general equation for a change in welfare measure in a discrete choice modelling framework is

$$\Delta E(\mathrm{CV}) = \frac{1}{\gamma} \int_{j=1}^{J} \mathrm{prob}_1(V_j^0, V_j^1) dV_j \tag{4.45}$$

given J number of alternatives in each state and changing the utility associated with alternative j from V^0 to V^1. Applying this formula it is assumed that marginal utility of income is independent of

income and prices, and that income effect is negligible (i.e. the compensated (Hicksian) demand curve and the Marshallian demand curve approximate each other, (Small & Rosen 1981). The price coefficient can be specified to depend upon sociodemographics such as household size, but cannot be specified to depend upon income. Train (2003) notes that the conditions for using (4.45) are actually less severe than they might first appear. The formula can be used for policy changes that change consumer surplus by small amounts per person relative to income, even though the marginal utility of income in reality varies with income (i.e. diminishing marginal utility of income). The 'multiple alternatives' approach takes the random component of utility into consideration and hence the distributional assumption of the error terms impact the estimation of the change in welfare. Assuming that the random component is iid extreme value distributed provides some attractive features, as the integral then takes a closed form and the welfare measure is easy to calculate:

$$E(\mathrm{CV}) = \frac{1}{\gamma}[\ln \sum_{j=1}^{J} \exp(V_j) + C \tag{4.46}$$

Recall that this term is the same as that derived for the nested logit – the *log-sum* term. E(CV) is the average consumer surplus in the subpopulation of people who have the same representative utility. The total consumer surplus in the population is calculated as the weighted sum of E(CV) over a sample of decision-makers, with the weights reflecting the numbers of people in the population who face the same representative utilities as the sampled person. The C denotes an unknown constant and represents the fact that the absolute level of utility cannot be measured. The constant is irrelevant from a policy perspective as only differences in welfare (utility) matter (Train 2003). For a difference in welfare, CVM becomes

$$\Delta E(\mathrm{CV}) = \frac{1}{\gamma}[\ln \sum_{j=1}^{J^0} \exp(V_j^0) - \ln \sum_{j=1}^{J^1} \exp(V_j^1) \tag{4.47}$$

As an example, suppose that a DCE intends to examine the welfare gain associated with entry into the market of an improved asthma treatment. It is assumed that there are four products currently in the market (Alt1-4), and that the estimated utility scores associated with each treatment (alternative) and the price coefficient are as follows:

Alternative	Utility score
Alt1	0.5
Alt2	0.4
Alt3	1.2
Alt4	0.9
Choosing no treatment	0.0
New treatment	2.3
Price coefficient ($)	-0.0057

In this case the expected CVM becomes

$$\Delta E(\mathrm{CV}) = \frac{1}{0.0057}[\ln(e^{.5} + e^{.4} + e^{1.2} + e^{.9} + e^{.8} + e^{.0} + e^{2.7}) - \ln(e^{.5} + e^{.4} + e^{1.2} + e^{.9} + e^{.8} + e^{.0})] = 122\$$$

It is easy to see that in the case in which only one alternative exists for each state, the 'multiple alternatives' approach collapses to the 'State of the world' approach, and that ΔCV > ΔE(CV) for multiple alternatives. One of the potential shortcomings of the 'multiple alternatives' approach is the specification of the number of alternatives to be included in the model. Von Haefen (2003) states that specifying the number of alternatives requires judgement by the researcher that is often arbitrary and lacking theoretical or intuitive appeal. Silva (2004) argues that *E*(CV) is a less robust estimate than CV, due to the fact that *E*(CV) requires consistent estimation of the choice probabilities, which generally depends on the correct specification. He also notes that *E*(CV) is very vulnerable to biases affecting choice probabilities, such as status quo bias. Silva warns, therefore, that caution is needed in interpreting such results.

Discussions of the application of the two approaches have recently been published in the health economic literature. This includes the article of Lancsar & Savage (2004) and the associated comments of Silva (2004) and Ryan (2004b). Lancsar & Savage (2004) argue that many health economists have incorrectly applied the 'State of the world' model in DCE settings, with the consequence that welfare measures have been overestimated. It is possible that many health economists have overlooked the 'multiple alternatives' model. Future health economic papers will show whether researchers have taken note of the recent discussion.

As illustrated earlier, application of the logit model for welfare estimation provides some special properties and this model is thus widely used. Small & Rosen (1981) have shown the formula for welfare measurement with the probit model – due to the normalization assumption, simulations are necessary. Estimation using the random parameters model is more complicated than shown for the MNL model, as some (or all) of the parameters are assumed to be random. Train (1998) outlined an estimation procedure for welfare measures for a random parameters logit model; for an example of an application of the procedure, see (Boxall et al. 2003). Extensions of traditional welfare measurement include development of a conditional welfare measure – a measure that incorporates the implications of an individual's observed choice and hence is suitable for DCE with multiple choice sets per respondent (von Haefen 2003), and incorporation of temporal dependence (Swait et al. 2004).

4.4.8 Derivation of standard errors of MRS and WTP

Specifying the standard errors of MRS (MWTP) and welfare estimates (WTP) is more complex as these are derived from a ratio of two uncertainty estimates (the coefficients). Standard errors are informative to estimate as it is often useful to specify the confidence interval of the given ratio. An approximate solution is to apply the following expression for the variance of the ratio of the estimates (the delta-method using Taylor's approximation series). Assuming a linear utility function, the variance of a MRS estimate is (Bateman et al. 2002):

(4.48) $$\operatorname{var}\left(\frac{\beta_j}{\beta_i}\right) = \left(\frac{\beta_j}{\beta_i}\right)^2 \left(\frac{\operatorname{var}(\beta_i)}{\beta_i^2} + \frac{\operatorname{var}(\beta_j)}{\beta_j^2} - \frac{2\operatorname{cov}(\beta_i, \beta_j)}{\beta_i \beta_j}\right)$$ [38]

Other techniques that are frequently applied are simulation methods such as the Krinsky-Robb method and boot trapping. The Krinsky-Robb method developed by (Krinsky & Robb 1986) involves randomly drawing the coefficients a given number of times (specified by the researcher) from the multivariate normal distribution of the coefficients and covariance matrix, and calculates the far equivalents for each of these draws. In the application of boot trapping, a number of new

[38] Calculation of part-worths standard errors in the random parameter model can be extended from the application of mean population estimates (as shown in the equation) so as to take into account all the information obtained from the model about heterogeneity and thus to calculate individual part-worth parameters. This is associated with non-trivial problems, however (Iraguen & Dios Ortuzar 2004)

data sets are created using the estimated residuals, and then the function is re-estimated. Although the two methods give similar results, the boot trapping procedure is more computationally burdensome than the Krinsky-Robb method (Carlsson et al. 2003).

When investigating methodological DCE issues, it might be appropriate to test for preference differences among sub-samples. For instance, a survey might be designed such that it is possible to test variation in perception related to the framing of the payment vehicle (keeping other elements of the questionnaire the same and randomly splitting the sample into subgroups, each of which receives a different variant of the questionnaire). Due to possible scaling differences among the sub-samples, one cannot compare the absolute parameter estimates of the two samples[39]. Instead, one can test for statistical differences in MRS and WTP as these are not subject to scaling, i.e. the scaling factor cancels out when coefficients are divided (Train 2003), e.g.

$$\mathrm{MRS}_{12} = \frac{\beta_1 \mu_{sample1}}{\beta_2 \mu_{sample2}} = \frac{\beta_1}{\beta_2}$$

Testing for differences in MRS/WTP is performed by *t*-ratio testing or, alternatively, by examining whether the point estimates (mean of the parameter vectors) of each sub-sample lie within the confidence interval of the parameter estimates of the other sub-samples.

[39] Simple comparison of parameter estimates is only possible under the assumption of equal scale among the samples.

5 Design of the DCE

The design of the DCE is very important as this influences how much information can be extracted. In general, the term 'design' refers to the science of planning in advance exactly which observations should be incorporated in the choice experiment in order to permit the best possible inference from the data (Louviere et al. 2000). The construction of the choice task must be based on the main elements that influence the choice in question. It should be determined whether there is any methodological issue that requires investigation - if so, it may be necessary to set up parallel DCE surveys which differ only in terms of the factor relevant to that issue. For example, it may be hypothesized that the value estimate is influenced by the inclusion/exclusion of an attribute or the framing of an attribute.

It has been – and still is – common practice to divide the design procedure into different stages, according to the different issues that need to be considered. This makes the process easier to manage and more consistent. Various researchers from a range of disciplines have developed and published design models (e.g. Green & Srinivasan 1978;Gustafsson et al. 2000;Louviere et al. 2000;Ryan 1996). These design models have many similarities, but differ on a number of issues such as the number and dimensions of the proposed stages. The design models should not be taken as definitive, strict rules but seen as guidance in the design process.

The author of this paper draws heavily on the design model proposed by Ryan (1996b). This is first and foremost because this model is favoured in the health economic literature (e.g. Carlsson & Martinsson 2003;Ratcliffe & Buxton 1999;Ryan 1996;Ryan 1999a). The design is divided into 5 stages, as illustrated below (Figure 5.1).

It was previously the practice to divide stage 4 into two stages: one in which the alternatives are created and one in which the alternatives are paired. Current access to computer programmes that can create and pair alternatives in one step has changed this, however. Furthermore, the previous descriptions of these stages referred to the possibility of using different choice procedures (e.g. whether to use DCE, ranking or rating); as DCE is by far the most widely used method today, this design question is of little relevance.

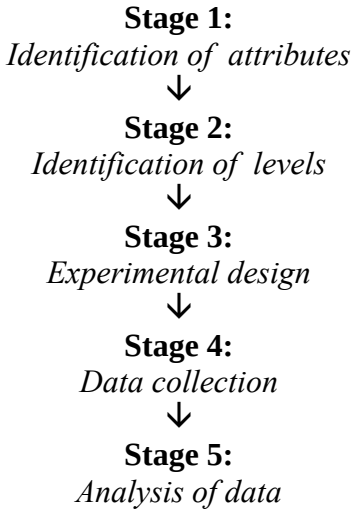

Figure 5.1: The design stages of a DCE

Before stage 1 of the design is embarked upon, it is crucial that the researcher has established the issue for investigation, i.e. the characteristics of the decision problem and the objectives of the study. In defining the decision problem, both the choice context and the economic value under consideration must be decided, as these greatly influence the design task. As an example, consider a situation where policy makers are contemplating the introduction of a new cancer treatment. The DCE analyst must clarify, for instance, whether only the use values associated with the service should be estimated, or whether non-use values and option values should also be estimated. Depending on the decision made, different attributes, levels and population groups may be specified. Besides this, it is always important that the choice context and the alternatives presented are understandable, and the respondents must be able to have a degree of confidence that the alternatives reflect actual possibilities (Bennett & Blamey 2001).

In the following sections, each design stage is discussed in more detail.

5.1 Stage 1: Identification of attributes

The first stage of the design involves defining the attributes of interest. While a number of ways exist to do this, no 'gold standard' has been identified. There is virtually no consensus about how to define the attributes, other then a consensus that it has to be done (Louviere 2000). Two issues in particular need to be considered when deciding which attributes will be included in a study. First, the attributes should be relevant to the requirements of the policy makers. Secondly, the

attributes need to be meaningful and important to the respondents (Bennett & Blamey 2001). In order to ensure that these requirements are satisfied it is important to obtain as much information as possible – and from different sources. These may include literature reviews, group discussions (e.g. focus groups), interviews with key persons such as policy makers, and expert opinion.

To be useful, an attribute should be both comprehensive and measurable. An attribute is comprehensive if, by knowing the level of an attribute in a particular situation, the respondent has a clear understanding of the extent to which the associated objective is achieved. An attribute is measurable if it is reasonable i) to obtain a probability distribution for each alternative over the possible levels of the attribute and ii) to asses the respondents' preferences for different possible levels of the attribute, for example in terms of a utility function. In line with this, Blamey et al. (2002) state that preference should be given to those attributes that are demand-relevant, policy-relevant and measurable. According to Keeney & Raiffa (1976), in order to enable decision making, it is important that the chosen attributes have the following properties:

- Completeness: The attributes cover all the important aspects of the issue in question
- Operational: That the attributes are meaningful
- Decomposable: That aspects of the evaluation can be broken down into parts of smaller dimensionality
- Non-redundancy: The attributes should be defined such that double counting of consequences is avoided
- Minimum size: It is desirable to keep the set of attributes as small as possible.

There are no general 'rules' for the number of attributes to be chosen, although there seems to be a consensus to have a maximum of eight attributes.

It is important to distinguish between attributes that are relevant and those that are irrelevant. An attribute is said to be relevant if ignoring its existence would change the conclusions, and irrelevant if ignoring its existence would not change the conclusions. Distinguishing relevant from irrelevant attributes is not always easy, however – one reason is that attributes that are demand-irrelevant under current levels of consumer awareness and involvement may become relevant and determinant under higher awareness (Bennett & Blamey 2001). Exclusion of important attributes will most likely result in biased estimates and inaccurate welfare measures, as the model to be estimated (and thus the utility values) is completely dependent on the attributes included in the experiment. Such

studies are dangerous if they remain undetected, as they provide only misleading information; the good in question is not truly specified, and the possibility of omitted variable bias has been introduced.

The researcher needs to be aware of problems related to two types of attributes: mutually dependent attributes and causally related attributes. Mutually dependent attributes are attributes that are dependent in some way on each other. For instance, in the evaluation of job opportunities it could seem appropriate to include an attribute for travel time to work and an attribute for the degree of flexibility to work at home. These two attributes are mutually dependent, however: if the respondent gets the opportunity to work at home one day at week, this will influence the disutility for travel time to work, as the total travel time is reduced. Mutual dependent attributes can be handled either by combining the two attributes into one, with the result that some information will be lost as the effects of the attributes cannot be separated (introducing correlation to the experimental design); or by defining one of the attributes in the introductory text to the choice task. Inclusion of one attribute in the introductory text means that, while both attributes are still defined, only one will appear in the choice sets. There is no reason why the attribute defined in the introductory text cannot be assigned different levels (sub-sampling). Excluding one of the attributes from the study is probably the simplest method to implement, but is problematic as the missing attribute would then be subject to the respondents' assumptions, introducing omitted variable bias.

Another common and related problem in the definition of attributes is that some attributes might appear to be causal to other attributes, i.e. causally related attributes. Blamey et al. (2002) investigate this issue in a study of habitat protection. They split their sample into two: one in which the causal as well as the effect attribute is present ('loss in area of unique ecosystems' and 'number of non-threatened species lost') and one in which only the effect attribute is present ('number of non-threatened species lost'). Interestingly, they find that the part-worth utilities have been repacked such that the overall welfare implications of a policy proposal are unchanged. The utility associated with the causal attribute is 'transferred' to the effect attribute when the causal attribute is absent (the utilities offset each other); the utility associated with the effect attribute changed by 34% when the ecosystem attribute was included. The authors consider the study result to be encouraging as far as the use of DCE for welfare estimation is concerned. However, they draw attention to the fact that causally related attribute subsets are identified at a preliminary design stage and addressed accordingly. They note that the challenge to the researcher is to handle the task such that reliability and validity is maximized.

5.2 Stage 2: Identification of levels

Stage 2 involves assigning levels to the identified attributes. One of the first steps is to determine the way in which levels are to be presented: either qualitatively or quantitatively. Moreover, it has to be decided whether the quantitative attributes should be presented in absolute or relative terms (for instance compared to status quo) (Bennett & Blamey 2001). A special issue relates to the inclusion of a risk attribute, as studies show that the framing of the risk attribute highly influences the respondents' choices and hence the estimated parameters. Ryan (1999a) describes three key success factors when choosing the levels for each attribute:

- The levels must be *plausibl*e to the respondents
- The levels must be *actionable* to the respondents
- The levels must be constructed so that the respondents are willing to make *trade-offs* between combinations of the attributes.

5.2.1 Level range

As noted above, the respondents need to be willing to make *trade-offs* between combinations of the attributes. This issue is important to address, as an improper level range can result in biased estimates. Too narrow or to wide a distance (interval) between the levels might result in the respondent considering the difference to be insignificant or significant resulting in dominated or dominating levels, thus leading to non-trading (seemingly dominant behaviour). The result may thus be insignificant or extreme estimated coefficients, respectively. It is important to keep the levels in mind when interpreting the results, therefore. An insignificant coefficient does not necessarily mean that the attribute is unimportant to the respondents; the correct interpretation is that the attribute did not influence the choices for the given levels.

As pointed out by Green & Srinivasan (1990), the levels have to be acceptable such that levels that will be dominated at any stage are avoided. This is not easy to determine, however, as it requires the researcher to have in-depth prior knowledge about which levels are acceptable to the respondents. Furthermore, the situation becomes much more complicated if there is considerable heterogeneity in preferences - which can usually be expected. The construction of attribute levels from the perspective of the 'average respondent' may prevent respondents with more extreme

preferences from making trade-offs between attributes. Ohler et al. (2000) reported that goodness-of-fit and heterogeneity was affected by changes in the outer level range, but that these range differences did not appear to have any (or at least little) effect on the coefficients or error variance. This result is supported by Ryan & Wordsworth (2000), who found that the difference in the inner range of some of the levels had insignificant effect for five out of six estimated coefficients; but nevertheless the welfare estimates were affected. Ryan & Wordsworth (2000) examined the sensitivity of the estimates to the level of attributes in a study of cervical screening programmes. They used two questionnaire variants that differed in the level range of three of the seven attributes – two qualitative attributes and the cost attribute. They found that WTP for attributes was significantly different for four of the five qualitative attributes[40], whereas the total WTP (the welfare estimate) for two hypothetical policy changes did not differ substantially. Whether the results from these two studies are generalisable to other contexts is not certain. Further research on this issue would be useful, in particular examination of the effect of level range on the variance of the unobserved effects. In this respect, Louviere (2001f) showed that the variance of the random component (response variability) was influenced by the attribute levels presented to the respondents.

Additional arguments exist for the observation of dominant preferences. This may be due to respondents having lexicographic preferences for one attribute, i.e. that they are unwilling to trade that attribute at all. If so, this will cause a violation of the continuity axiom and a departure from the use of compensatory decision-making. Dominant preferences may also be a consequence of cognitive inabilities or bounded rationality (due to the complexity of the choice task) that result in the use of heuristics (such as fast and frugal heuristics) in order to answer the questions (e.g. Scott 2002). This issue is discussed more fully in chapter 6.

5.2.2 Attribute-effect

It has been shown that the number of levels in a choice experiment influences the significance of the attribute. This is known as the *attribute-effect* and refers to the situation where an increase in the number of levels for an attribute – without changing the upper and lower levels – causes the attribute to become relatively more significant. Imagine a cost attribute with three levels (150 kr., 200 kr. and 250 kr.) and a cost attribute with five levels (150 kr., 175 kr., 200 kr., 225 kr. and 250

[40] Using a *Z* statistic

kr.). Even though the attributes cover the same monetary interval, the attribute with five levels will have a bigger impact on the model compared to the attribute with three levels. Ratcliffe & Longworth (2002) investigated this issue in a study of alternative modes of intrapartum care and found that the respondents placed more importance upon attributes with a high number of levels. This attribute-effect can bias the result, and affect the modelling. One way to minimize the problem is to assign the same number of levels to every attribute (Curry 1997). This will neither be practical or desirable in most cases, however. The issue is of special importance in relation to the cost attribute, as this attribute is used as the numerator for WTP estimation and is often assigned more levels than other attributes (such as 6-8 levels). Figure 5.2 illustrates attribute-effect and level range differences.

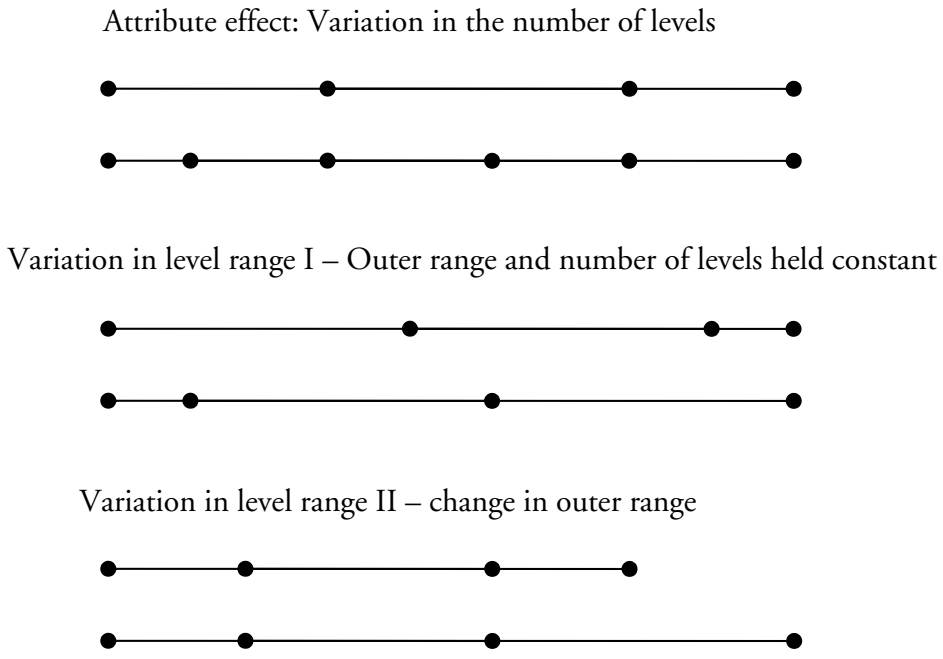

Figure 5.2: Illustration of attribute-effect and level range.

5.3 Inclusion of a cost attribute

If the researcher is only interested in the implied ranking of attributes, or in their part-worths relative to each other, then cost does not need to be one of the attributes. Many studies do not incorporate a cost attribute and the results of such analyses are just as meaningful as results from

studies that do include a cost attribute. It might even be argued that studies that exclude a cost attribute are advantageous as they avoid the well-known problems that arise from inclusion of a cost attribute and estimation of WTP values. On the other hand, the incorporation of a cost attribute increases the capabilities of a DCE. If the aim is to estimate a welfare-theoretic (and hence comparable) estimate of benefits, say, for use in CBA, then a cost term will need to be included in the design as a proxy for marginal utility of income (money). Bryan et al. (1998) examined the effect of including a cost attribute in their evaluation of treatment for knee injuries by conducting two studies: one with and one without a cost attribute. They found that the study including the cost attribute had more missing data, which might indicate a possible resistance to the inclusion of a cost attribute. Researchers argue that some of the potential problems related to stated preference methods, such as protest bids, are limited in the context of DCE (relative to CVM), due to the fact that respondents are not asked their WTP directly; hence influence of the cost attribute on response patterns is reduced.

The way in which cost is expressed is a design issue that creates very similar problems to the choice of payment vehicle in CVM. Clearly, the cost tag needs to be credible and realistic, and ideally should also minimize incentives for strategic behaviour. Bateman et al. (2002) discuss respondents' incentives to give truthful responses in SP settings. They argue that it is important to use an incentive compatible elicitation procedure, i.e. a procedure in which the questions are formulated in such a way that it is in each respondent's interest to give truthful answers. Thus the question of incentive compatibility is highly relevant in choice of payment vehicle. Bennett & Blamey (2001) are correct in arguing that the framing of the task and how it affects individuals' perceptions of their payment obligations and the impact of their choices on the potential for supplying the good must be examined. If, for instance, individuals feel that they would never be charged the fee or tax increase listed in the choice scenario, then they may behave strategically when responding to the choice questions. Regardless of which type of payment vehicle is included, there will be some kind of concern associated with it. For example, using tax as a payment vehicle involves equity concerns, whereas donation involves concerns of free-riding and 'purchase of moral satisfaction' (Kahneman & Knetsch 1992). Green et al. (1998a) noted from the CVM literature that responses are influenced by the payment vehicle used. They pointed out that such effects may arise from incentive effects of the free-rider variety, or from the concerns of subjects about distributional implications and fairness.

The inclusion of cost might be additionally problematic in health-related DCEs. The use of cost to estimate WTP raises questions about the definition of the cost attribute in a collectively funded health care system, where the inclusion of cost might result in the scenarios being considered unrealistic and immoral, possibly resulting in protest bids[41]. The choice of payment vehicle thus has to be considered very carefully in these types of studies. Most health care services in Denmark are free at the point of consumption; approximately 80% of the health expenditure is paid by the government through taxes, while the remaining 20% is financed through user payment. Among the health services that are partly financed by users (by means of a co-payment system) are pharmaceuticals and dental services. When evaluating a health care service such as a pharmaceutical, therefore, the payment vehicle can fairly easily be identified as an 'out-of-pocket' price. In the evaluation of other health care services and interventions such as hospital treatment, however, other payment vehicles need to be considered, such as tax payment.

Boardman et al. (2001) reasoned that preferences for payment vehicles might differ and hence these preferences should be taken into consideration in the evaluation. Skjoldborg & Gyrd-Hansen (2003) examined the effect of different payment vehicles (tax payment versus user's charge) in relation to choice of hospital and preferences for the Danish health care system in general. The results indicated that respondents reacted differently to the payment vehicles and the negative utility associated with paying per se was greater when the context used was user charges. It was found that a 1000 kr. increase in tax payment was not considered equivalent to a 1000 kr. 'out-of-pocket' payment. If this result can be generalized to other settings, it questions the external validity of the DCE and calls for greater consideration of the choice of payment vehicle. The researcher also needs to decide upon the payment duration: should it be identified as a weekly, monthly or perhaps yearly payment? If WTP values prove to vary depending on the chosen period, and that marginal WTP decreases relatively as the period increases, due to the larger amount described, it argues in favour of applying what is considered common practice, or alternatively the application of long durations, as these underestimate WTP values that can then be considered conservative estimates.

As mentioned above, the setting of appropriate levels for the cost attribute (both in number and range) is an important consideration in the estimation of WTP. Choosing an inappropriate cost interval may generate over- or underestimated WTP values and hence misleading results. For instance, setting to low levels may result in non-trading of the cost attribute and hence a higher probability for choosing a given choice alternative, all other things being equal. This would imply a

[41] Respondents refuse to answer the question because they have a moral objection to being asked about their WTP (Mitchell & Carson 1989). See also chapter 6

higher positive utility for the alternative and less disutility for paying for it - and thus an overestimation of WTP.

5.4 Stage 3: Experimental design

> *"The way the alternatives' levels are set and structured into choice sets in known as the 'experimental design'." (Bennett & Blamey 2001, pp. 57).*

In this stage of a DCE, the hypothetical choice sets are designed – including the formation and pairing of alternatives. Various methods can be used to design and reduce (if required) the number of choice sets to be included in the questionnaire. In this regard, one of the crucial objectives of the experimental design is to create the DCE in such a way that the number of alternatives is minimized while being able to infer utilities for all possible alternatives – which implies keeping the choice task simple to the respondents and at the same time being able to extract all the necessary information from the choices (i.e. securing a high degree of design efficiency). A designed experiment is therefore a way of manipulating attributes and their levels to permit rigorous testing of certain hypotheses of interest (Louviere et al. 2000). Louviere et al. (2000) further state that experimental design for non-linear models, such as the design for DCE choice models, is in its infancy, although advances are being made. This issue is discussed more fully in chapter 4.

Experimental design is one of the foundations for a successful DCE and requires considerable thought. An overview of the general experimental design theory is given in the following sections.

5.4.1 Factorial design

As mentioned earlier, the number of possible alternatives increases exponentially when the number of attributes and levels increases (# alternatives = $\#levels^{\#attributes}$). The overall question is, therefore: Is it possible to present the respondents with all possible alternatives or is it necessary to reduce the number presented to them? This of course depends on the number of attributes and levels and the number of alternatives (choice sets) which respondents can cope with. First of all, it is important to reduce the number of attributes and level to a minimum – this is considered in stage 1 and stage 2. Next, methods exist that can reduce the number of alternatives included in the questionnaire while keeping the statistical properties of the design. One of the objectives of a DCE is to calculate the MRS between different attributes - and in particular welfare estimates if cost is

included. The accuracy of the estimated parameters is of *vital importance* and depends on the efficiency of the preferences that are elicited (Louviere et al. 2000).

Many different design methods exist, with varying difficulties and reliability. The most popular way of combining the levels of the attributes is the use of *factorial design*; this paper will therefore be restricted to reviewing this type of design. Factorial design has very attractive statistical properties from the standpoint of estimating the parameters of models that test hypotheses. The questionnaire is designed so that each level of each attribute is combined with every level of all other attributes. A factorial design is simply the factorial enumeration of all possible combinations of attribute levels – for instance if we have 3 attributes, each with 2 levels, then the factorial will be 2^3= 12, implying 12 possible combinations of attribute levels, constituting 12 alternatives.

Factorial design can be divided into 'full factorial' design and 'fractional factorial' design.

Full factorial design

Full factorial design refers to a design in which all the possible alternatives are represented – in the above example this would mean that the respondents are given all 12 alternatives or, alternatively, that the survey population is split into sub-groups, where each sub-group answers a sub-sample of the choice sets. Full factorial design has very attractive statistical properties as it guarantees that all attribute effects of interest are truly independent (i.e. attributes are independent by design). However, full factorial design is only a real possibility for small experiments that involve a limited number of either attributes or levels, or alternative a highly blocked design. A choice experiment often involves 4-10 attributes, each with 2-4 levels. Consider a choice experiment with 5 attributes each with 4 levels; then the full factorial design results in 5^4 = 625 alternatives, which is a considerable number. With a large number of attributes and levels, therefore, it may be necessary to reduce the size of the design. This can be done by the use of fractional factorial designs (Louviere et al. 2000).

Fractional factorial design

Fractional factorial design involves a selection or a subset (a fraction) of the original full factorial design, in which the properties of the full factorial design are maintained in the best way possible, such that the effects of interest can be estimated as efficiently as possible. Instead of random selection, statisticians have developed sampling methods that lead to practical and

manageable designs with specific statistical properties. However, it is important to be aware that all fractional designs involve some loss of statistical information. This loss of information can sometimes be significant, as fractional factorial designs limit the ability to take higher order effects[42] into account, i.e. interactions between two or more attributes (Louviere et al. 2000).

Most studies use what is called a main effect fractional factorial design, in which it is assumed that interactions among attributes are insignificant in all two-way and higher order interactions. As mentioned in chapter 4, such interactions are included to take into account the possibility that respondent's preferences for one attribute depend on the level of another attribute. The assumption of no interaction requires the belief that two-way and higher interactions do not bias the result in any way. If the omitted interaction effects are significant, then the results estimated by such a fraction will be biased and the exact nature or degree of the bias cannot be known in advance due to the unobserved interactions (Zwerina et al. 1996). On the other hand, Louviere et al. (2000) argue that the exclusion of interaction effects does not necessarily lead to biased result, because:

- Main effects typically account for 70% - 90% of the explained variance
- Two-way interactions typically account for 5% - 15% of the explained variance
- Higher-order interactions account for the remaining explained variance.

Even if interactions are significant, therefore, they rarely account for much of the explained variance and hence will not affect design efficiency vitally. This lead Louviere et al. (2000) to argue that, in the light of the difficulties with explanatory power that arise from the inclusion of interaction terms, it seems fair in most cases to exclude the interaction terms.

5.4.2 Ensuring high design efficiency

In the early days of DCE, the experimental design (such as selection and pairing of the alternatives) was done unsystematically – for example, by making random draws in which a fraction of the alternatives was chosen randomly. However, this way of designing DCEs reduces

[42] Inclusion of interactions depends on the degrees of freedom; interaction effects to be considered should thus be pre-specified.

statistical efficiency[43] was not necessarily high – and in the worst case, the DCE turned out to be useless. For instance, imagine a pairing of alternatives in which one attribute always increases as another attribute increases. In this situation of collinear attributes, it is impossible to isolate the effect of each attribute on overall utility. Tools have been developed to deal with this problem, in order to ensure the statistical properties of the design.

When using a full factorial design, all the possible alternatives are in use. This is not enough to ensure high design efficiency, however. What appears to be important is the pairing of the alternatives into choice sets. The pairing of alternatives needs to be made in such a way that the *differences*[44] in attribute levels for each choice set are not multi-correlated. Moreover, if each respondent is not presented with all the choice sets but only a fraction of them – e.g. if the respondents are split into two groups where each group is given half the choice sets – then it is important that the choice sets are spilt into two blocks that retain their statistical properties. These two considerations also apply to fractional factorial designs. In addition, as only a fraction of the total possible alternatives is to be presented to the respondents in fractional factorial design, the selection of alternatives also needs to considered in the light of design efficiency. See Table 5.1.

Table 5.1:The steps of experimental design in which design efficiency has to be taken into consideration

Full factorial design	**Fractional factorial design**
• Pairing of alternatives into choice sets • Block design	• Selection of alternatives • Pairing of alternatives into choice sets • Block design

To ensure – in theory – maximum statistical efficiency in choice design (i.e. the extraction of maximum information from the choice task), there are some properties that have to be jointly satisfied. Together, these principles are called design efficiency, also termed *D-efficiency*. *D*-efficiency relates to the design matrix in such a way that efficiency is maximized when the size of the covariance matrix of the estimated parameters is minimized. Huber & Zwerina (1996) identifies

[43] Efficiency: Goodness of an experimental design. Efficiency can be quantified as a function of the variances and covariances of the parameter estimates. Efficiency increases as the variance decreases

[44] Remember that probabilities (utilities) are estimated as the difference in attribute utilities for each alternative, $\Delta V=\beta\Delta X$

four principles for an efficient design based on a non-linear model[45]. To optimize *D*-efficiency, four principles need to be considered simultaneously. Improving any principle, holding the others constant, improves efficiency. . In many cases it is impossible to create a design that satisfies all four principles, as some of the principles might conflict with each other. The four principles are[46]:

- Level balance
- Orthogonality
- Minimal overlap
- Utility balance

Level balance

Level balance simply means that the levels of an attribute occur with equal frequency in the design, e.g. each level of a four-level attribute should occur in precisely one-fourth of the included alternatives. This ensures that all levels are weighted equally in the trade-off options that the respondent faces (Huber & Zwerina 1996). If all alternatives are given to each respondent, then level balance is already ensured. However, with the use of a block design or a fractional factorial design, level balance needs to be taken into consideration in order to optimize efficiency.

Orthogonality

Orthogonality can with reason be considered the most important aspect of *D*-efficiency. Orthogonality is satisfied when the joint occurrence of any two levels of different attributes appears in profiles with frequencies equal to the product of their marginal frequencies (Huber & Zwerina 1996). Orthogonality is thus satisfied when the difference in the levels of each attribute varies independently over choice sets, meaning that the levels of the attributes vary in a criss-cross manner. As 'pure' optimal orthogonal designs are only available for a very small number of very specific problems, the primary purpose is to optimize the design as best one can, i.e. make it as efficient as possible, by minimising multicollinearity. (Kuhfeld et al. 1994). A high degree of multicollinearity will result in a design in which unique estimates of the parameters cannot be

[45] The model is non-linear due to the discrete nature of the dependent variable
[46] There are a number of other statistical designs that consider certain of the four principles, including cyclic designs and orthogonal designs (Carlsson & Martinsson 2003)

obtained, making it impossible to draw any statistical inferences, i.e. hypothesis testing, from the sample.

Minimal overlap

Minimal overlap relates to the statistical properties when pairing the alternatives. A design has minimal overlap when a level does not repeat itself in a choice set. In order to optimize orthogonality of the level differences, the scenarios are matched to ensure minimal overlap (i.e. optimal orthogonality ensures minimal overlap). Minimal overlap is important in choice designs, because the contrast between attribute levels is only meaningful as differences within a choice set. Minimal overlap ensures that the probability of an attribute level repeating itself in each choice set is as small as possible, and thus maximizes the information obtainable from the choice sets. The cost of violating this criterion can be seen most clearly when the levels of one attribute are the same across all alternatives within a choice set. The choice set then provides no information on the value of the attribute in question (Huber & Zwerina 1996).

Utility balance

Utility balance is a relatively new approach that has not yet had much attention. It can be rather problematic to incorporate in the design as it demands a priori knowledge of respondents' preferences. A choice set is utility balanced when the utilities of alternatives within each choice set are approximately equal. To achieve this, the researcher needs to take the utility weights of the attributes into account when designing the DCE. The rationale for this principle is to ensure that respondents are actually trading. The efficiency gain arises because choices between alternatives that have similar utility provide better information about the coefficients. This means that two alternatives that differ in their levels but have approximately the same utility are more likely to ensure that the respondents are placed in a situation in which they are forced to trade. Application of the utility balance concept thus implies that the impact on choices of small differences in utility is registered resulting in more precise parameter estimates.

What makes this principle troublesome, however, is that it requires prior estimates of the coefficients. There are several ways to generate useful sets of prior estimates. The most used method is to conduct a small pilot study to generate tentative estimates. Carlsson & Martinsson (2003) and Huber & Zwerina (1996) have shown that the incorporation of the utility balance principle increases efficiency of the DCE.

Two considerations need to be given to the principle of utility balance. One is that utility balance is estimated for the average respondent – meaning that respondents with more extreme preferences are not forced to make trade-offs (the utility balanced design thus works best for respondents with homogenous preferences). The other consideration is that a utility balanced design might be cognitively demanding for the respondents, as each choice is difficult to make – hence increasing the likelihood of non-response and irrational responses such the use of heuristics in the decision-making process. This issue is further discussed in the next section.

Nowadays there are computer programmes that can operate with *D*-efficiency, thus facilitating the design process considerably. SAS is the most widely applied package for experimental designs given that there exists a range of macros designed for different choice task purposes (see Kuhfeld 2003). On the basis of the researcher's choice restrictions (such as the number of attributes and levels, the number of alternatives in each choice set, and the number of choice sets), the computer generates an experimental design that optimizes *D*-efficiency and makes it possible to extract the maximum amount of feasible information. Zwerina et al. (1996) note that the use of computer to directly minimize *D*-error may only approximately satisfied the four principles in the designs, however, the designs generally being more efficient than those built directly from the principles. See Carlsson & Martinsson (2003) for a discussion of design techniques for DCEs in health economic applications.

Researchers have come a long way in recognising the necessity of a high-quality experimental design. As with many DCE issues, however, experimental design is under continuous development. Hopefully, we will be able to control this aspect of the design process more satisfactorily in the future.

5.5 Complexity versus completeness - The extent to which design influences cognitive demand

It is important to keep in mind that experimental design efficiency not is the only important issue to be considered in the design process. Louviere et al. (2000) state that the design objective should also include considerations related to:

- **Identification of the utility function:** theoretical considerations need to be given as to whether it is appropriate to assume the particular form of the utility function - most often a linear additive utility function
- **Precision:** more precise estimates have smaller confidence intervals and hence greater statistical efficiency
- **Market realism:** the degree to which the DCE matches the actual decision environment faced by respondents; logically, the closer the experiment resembles the actual market, the higher the content validity.

The consideration of market realism relates to the fact that the use of statistical tools to reduce the number of scenarios to a manageable level raises the possibility that some of the scenarios presented by the statistical design may be unrealistic. This in turn raises the question of whether the statistical design should be compromised for realism, or realism compromised for statistical reasons (Ryan 1999a). Kuhfeld (2003) argues that the loss of statistical significance is not as distinct as initially assumed as long as the researcher tries to design the attributes and the levels in an optimal way. Moreover, Louviere et al. (2000) note that orthogonality not is the primary goal in design creation. It is a secondary goal, associated with the primary goal of minimising the variances of the parameter estimates. They consider the degree of orthogonality to be an important consideration, but note that other factors should not be ignored. In line with this, Bateman et al. (2002) report that although orthogonality is a desirable property in a choice task design, there are reasons why we may depart from it in practice – including plausibility and realism. They conclude that departure from orthogonality is best left to the researcher to decide but should always be justified, with supporting analysis of the statistical properties of the design. A final objective mentioned by Louviere et al. (2000) is:

- **Cognitive complexity:** the degree of task complexity and difficulty arising from the experiment. There is little consensus, and even less empirical data, regarding optimum levels of complexity.

This last objective warrants further discussion as this is a crucial consideration. Not only does cognitive complexity relate to the experimental design (Stage 3), but it also relates to Stages 1 and 2, and indirectly to the theoretical assumption of human behaviour. The design of a choice

experiment typically involves trade-offs between completeness and complexity. As the number of attributes and levels increases, the number of possible alternatives increases exponentially and the cognitive demands of fully evaluating the scenarios also increase dramatically; the extent to which this happens in practice remains unknown, however. There is no golden rule about the number of attributes and levels, although it appears to be the norm in the literature that 4-8 attributes are included in the study (Curry 1997;Ryan & Gerard 2003) and a norm to keep the number of levels to a minimum. Regarding the number of alternatives per choice set, there has been a tradition within the field of health economics to include two alternatives (this is starting to change, however), whereas in environmental economics the tradition has been three alternatives (including an 'opt-out' alternative). The number of levels and attributes that can be applied without causing too much trouble depends to a certain extent on the study's objective and mode of presentation, as well as the sample characteristics, but most importantly on respondents' familiarity to the subject matter and the design chosen, i.e. the number of choice sets and the number of alternatives per choice set that are presented to the respondents. The choice sets contain much information that must be assimilated by respondents and then acted upon. If the amount of information exceeds a respondent's cognitive ability, the respondent may refuse to answer any questions or, even more seriously, the answers given may not reflect true preferences either due to random answers or due to the use of decision-making short cuts or heuristics. This argues for the use of a simple choice task. On the other hand, however, there also needs to be sufficient variation in the choice sets and enough choice sets so as to establish the statistical properties necessary to estimate robust utility values (Bennett & Blamey 2001). Swait & Adamowicz (2001b) estimated an inverted U-shape relationship between choice complexity, measured by the number of choice sets, and variance of the latent utility function. Hanley et al. (2002) found that increasing the number of choice sets from four to eight had an insignificant effect on the value estimates, which indicated that this type of design decision has only a small impact on the elicited preference structure. Maddala et al. (2003) examined the design efficiency – complexity trade-off by studying the effect of increased overlap of attribute levels in the choice set. Their results were ambiguous. Interestingly, they did not find that increasing the amount of overlap resulted in significant improvement (such as higher consistency, less dominance, less perceived difficulty and less fatigue effect). Instead they found measurable differences in preferences without, however, being able to determine the cause of the differences found. In general, more research is needed that examines the relationship between task complexity and design completeness.

5.6 Stage 4: Questionnaire construction and data collection

Stage 4 involves all the remaining issues that need to be considered before the questionnaire is presented to the respondents.

5.6.1 Inclusion of an additional alternative: the opt-out/status-quo

In many cases it is appropriate to include an additional and fixed alternative (besides the ones chosen on efficiency grounds) in the choice set: an option not to choose any of the alternatives in the choice set – often referred to as an 'opt-out', 'neither', 'non-participating' or 'status-quo' (current) alternative[47]. The omission of such a non-choice option either forces the respondent to choose between alternatives not deemed important to him, i.e. alternatives that possess negative utility (implying misinterpretation and bias), or causes non-participation. It is therefore useful to include a non-choice alternative in situations where it is unrealistic that the good is consumed with certainty (refer to the earlier discussion of welfare measurement). Adamowicz et al. (1998) state that when dealing with expensive, risky new alternatives or technologies, there is a possibility that high levels of non-choosing will be exhibited. They argue that in such cases it is important to recognize that one wants to model not only the probabilities that consumers choose something, but also the probability that they will choose nothing, i.e. it becomes possible to allow for non-demanders. One should design DCEs to allow for the observations and modelling of non-choice as it is such an obvious element of real market behaviour. In line with this, Bateman et al. (2002) argue that the exclusion of an opt-out/status quo option is a violation of the underlying welfare measures. If the choice set does not include an opt-out alternative, then a non-zero value is implied in the estimated likelihood function for people who would not choose one of the alternatives. In general, this serves to bias the welfare estimates upward (Boyle et al. 2001). The inclusion of a non-choice option in the choice set makes it possible to estimate the value (i.e. welfare measure) of choosing one alternative compared to doing nothing – and not just the welfare change between two alternatives, which is all that is otherwise possible[48]. There are potential problems with the inclusion of such an option, however. If a respondent feels that the choice task is cognitively demanding, it will be tempting to simplify it by checking the non-choice option, simply to prevent

[47] Two important consequences of having a fixed alternative in the choice set are the need for including an alternative specific constant in the analysis and code dummy variables by effect coding (such that the dummy variables do not infer with the coefficient of the constant). It should also be noted that the inclusion of an extra alternative does not change the properties of the experimental design efficiency

[48] Instead of including an extra status-quo alternative, another possibility is to select the attributes and levels so that it is feasible to estimate both hypothetical and present alternatives.

making difficult choices -and not because it provides the highest utility among the alternatives - hence applying heuristics. A related problem is status-quo bias (loss-aversion). Studies show that individuals have a tendency to choose what they know even though it may seem irrational from a theoretical sense, as it provides less utility (e.g. Salkeld et al. 2000). Furthermore, it can be difficult for both the researcher and the respondent to know which levels are associated with the non-choice option, hence causing econometric and interpretation difficulties. This is especially the case if there is an additional (status-quo) alternative that differs among the respondents. Bennett & Blamey (2001) dealt with this issue in a study of saltwater fishing sites by asking respondents to indicate their status-quo (the specific site). This enabled the researchers to include the attribute information from the site in the choice task, hence increasing the amount of information obtained.

There is general consensus within environmental economics to add a non-choice option to the choice set when relevant (e.g. Adamowicz et al. 1998). In health care, however, the attempt to include such an option has been sporadic – and even when included in the questionnaire, the non-choice alternative might be excluded from the regression analysis (Ryan & Skåtun 2004). One reason for this might be that environmental economics is more advanced then health economics in the application of DCEs. It might also be argued that there are fewer appropriate situations for inclusion of a non-choice option in health economics - DCE has in many cases been applied to the estimation of recreational and similar values, in which the option of visiting the site is by nature optional (e.g.Hanley et al. 2002). Econometrically, one way of dealing with the additional fixed non-choice alternative is by use of the nested logit model – as described in section 4.4.3 – as the nested logit has the property of dividing the answers into categories.

Another issue relating to the inclusion of a fixed alternative is the option of including an 'indifference option' in the choice set, implying the identification of responder utility indifference – similar to ties in a rating choice experiment. A choice set can be designed in such a way that it just reaches the respondent's indifference point. Bryan et al. (1998) addressed this issue. In their design there was no 'indifference option' available to the respondents, however they reported that some respondents expressed their indifference for a specific choice set – for instance by ticking in the middle of the two boxes. The standard procedure nowadays is to omit such observations from the analysis, although they might contain valuable information. The same problems occur here as with the inclusion of a 'neither/opt-out' option, however: Can we trust these answers? Or are they a symptom of lack of preferences (i.e. violation of the completeness axiom), an indication of bounded

rationality, or the adoption of heuristics? Moreover, the indifference option raises an additional problem: How should such responses be interpreted and handled econometrically? Although these extra choice opportunities no doubt contain valuable information, consideration needs to be given as to whether the disadvantages outweigh the advantages.

Special consideration needs to be given in the situation in which a binary 'yes-no' design is chosen, see Figure 5.3. Imagine that different cancer screening programmes are to be evaluated. In a binary choice setting with the ability to opt-out, this implies a design in which one alternative is a generic screening alternative and the other alternative an opt-out alternative. When applying such a design, the researcher has to be sure that the respondents will trade-off screening versus no screening, and not always choose one or the other. In the case of screening, respondents might have a prejudiced attitude towards screening (no matter what the attribute level). If this is the case, the binary task will not provide the researcher with any information as trading has not taken place at the individual level (even though it might appear so at the aggregated level). The only information obtained in such a situation is the participation rate. No information is obtained regarding the relative weighting of the screening alternatives. The higher the degree of prejudice, the more problems there will be in applying such a method. Caution needs to be taken, therefore, when using a binary 'yes/no' design.

Much of the above discussion is also applicable for the inclusion of 'non-generic' alternatives such as labelled alternatives (e.g. bus versus car or payment versus no payment) considered in sections 4.4.3 and 4.4.4.

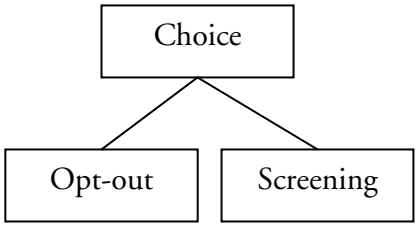

Figure 5.3: A binary 'yes/no' DCE

5.6.2 The introductory text

Before the respondents answer the DCE, it is important that they are first introduced to the task – for instance in the form of an introductory letter or preface. Two issues are especially

important to deal with. Firstly, the study objectives need to be specified - including the reasons for the choice of respondent group and the importance of the respondents' participation. Secondly, an introduction to the choice sets has to be provided. This should include the approximate time the task will require, assurance of confidential responses and, importantly, an explanation of the DCE task and, if necessary, a detailed description of the chosen attributes (Bennett & Blamey 2001). It has been argued that there might be a validity problem with the first choice sets answered, as respondents need some 'warm-up questions' in order to comprehend the rationale behind the choice task. Excluding, for instance, the first three choice sets in a questionnaire necessary implies, however, that three more choice sets are needed in order not to waste any information. It might be an advantage, therefore, to include an example choice set in the introduction so as to present the technique to the respondents before the beginning of the actual choice task.

5.6.3 Presentation of the choice sets

At the beginning of the choice task, a description is given of the overall choice scenario – the pre-choice text. This description places the choices into the context in which the respondents should make their decisions. In their study of out-of-hour primary care provided by general practitioners, Scott et al. (2003) used the following description (Box 5.1):

Box 5.1. An example of a pre-choice text

"Imagine that during the night, your child is short of breath, wheezing and coughing and that you decide to call a doctor. You have several options about the care you receive. These differ according to *whom* your child sees, where they are seen, the time it takes between making the telephone call and receiving treatment, and whether the doctor seems to listen to what you have to say.

For each question below, you are asked to choose which type of consultation you would prefer for your child during the night (Consultation A or Consultation B)."

Source: (Scott et al. 2003)

The framing of the presentation is very important. Bennett & Blamey (2001) argue that respondents to any stated preference questionnaire must be made aware that the good under consideration is embedded in an array of substitute and complementary goods, and that it might also be appropriate to remind the respondents of their budget constraint and other ways in which the money can be used. Furthermore, if the questionnaire deals with public funds, it is important that

the frame makes respondents aware of competing demands for public funds, i.e. the opportunity costs.

The questionnaire must be constructed in such a manner that there is an expectation that the information provided by the respondents will be used in some fashion for making decisions. If the respondents view the process as entirely hypothetical then their responses will not be meaningful in any economic sense (Bennett & Blamey 2001).

5.6.4 Inclusion of a validity test in the choice task

If the researcher wants to test for biases and the validity of the experiment in addition to what is possible from the chosen alternatives, it becomes necessary to include an extra choice set(s). For instance, if the researcher wants to test for rationality, a consistency test is needed. One of the simpler consistency tests involves the inclusion of a choice set in which one alternative unquestionably dominates the other(s) on all attributes[49]. 'Incorrect' responses can either be interpreted as a result of irrational respondents, a lack of understanding of the choice task, or a simple mistake on the part of the respondent. In some situations it might not be possible to define dominating alternatives, as no level is clearly preferable to another. In this case it is possible to include the same choice set twice in order to examine the consistency of preferences – it is assumed that rational respondents will answer the two questions uniformly. Another possibility is to include such choice sets that make it possible to test for rationality by testing for possible violations of the transitivity axiom: If $A \succ B$ and $B \succ C$ then $A \succ C$.

Other possibilities exist to test for biases and validation. These include tests related to:

- Framing – varying the framing of the questionnaire for different subgroups of the sample in order to test for a framing effect
- Ordering – varying the order of attributes, choice sets etc. for different subgroups of the sample in order to test for an ordering effect
- Level range – varying the level range for different subgroups of the sample in order to test for level effects
- Overlap – varying the amount of level overlap in a choice set for different subgroups of the sample in order to test for the effect of overlap on response variability and design efficiency.

[49] Normally such a choice set will not be generated using computer experimental design as this is not very efficient.

5.6.5 'Uncertainty question' and 'cheap talk'

One of the most criticized aspects of stated preference methods is that they are hypothetical in nature and hence suffer from hypothetical bias. No matter how individuals answer in such surveys, their responses have no consequences for them – they are just 'pretending' to buy the good, i.e. stating their value of a particular good. The hypothetical nature of stated preference methods implies that the methods are sensitive to overstatements of the amount individuals are willing to pay. This problem has been recognized for some time in the contingent valuation literature, and ways to deal with the problem include ex post correction by calibration and the inclusion of a request that respondents consider their budgetary constraints. One of the most widely applied calibration techniques in CVM studies is certainty calibration. Here, the WTP question is followed by a question which asks how sure respondents are about their response to the valuation question. This is usually accompanied by a 10-point scale for the level of certainty (Champ & Bishop 2001). A common application of the certainty scale is to treat positive answers as 'yes' only when certainty levels are at least 8 on a 10-point scale, where 10 indicates 'very certain'. Although results are mixed, it appears that the procedure may be capable of reducing hypothetical bias (Samnaliev et al. 2003).

Cummings & Taylor (1999) introduced a new method of providing bias correction into CVM studies – the 'cheap talk' approach. Lusk (2003) argues that the cheap talk approach has the advantage that it is a general approach, providing an ex ante bias correction that can be applied in any evaluation. The idea behind cheap talk is simply to make the respondents aware of the potential problem of overstating the payment bids prior to answering the valuation questions. This can be done in many ways, with the general idea being that respondents are faced with a discussion of the hypothetical bias problem – what hypothetical bias is and why it might occur (Cummings & Taylor 1999). Studies have shown that cheap talk effectively reduces WTP – making hypothetical valuation questions indistinguishable from responses to valuation questions involving actual payments. In particular, it seems that cheap talk has the property of reducing WTP to unknowledgeable respondents while keeping knowledgeable respondents' WTP unaffected (e.g. List & Gallet 2001). Lusk (2003) points out that many questions remain unanswered regarding the application of cheap talk, such as the interaction between knowledge and experience of the respondents, payment vehicle, and cheap talk effectiveness. Aadland & Caplan (2003) examine the length and wording of the cheap talk script and find these factors to influence the WTP estimates. They therefore suggest caution in using cheap talk script.

Box 5.2 below provides examples of an uncertainty question and a cheap talk script.

Due to the fact that DCE more closely mirror actual consumer purchasing situations than CVM, it has been hypothesized that DCE is less prone to the prevalence of hypothetical bias in WTP estimates, hence having better incentive compatibility properties (Lusk & Schroeder 2004). However, further research that examines whether DCE is subject to hypothetical bias is necessary to determine the validity of valuation method and potential advantages compared to CVM. There are only few studies that test for hypothetical bias in DCE settings, and to the author's knowledge the effect of uncertainty question and cheap talk have not yet been examined in any health economic settings. Recently, Carlsson et al. (2004) have tested the inclusion of the cheap talk. By comparing to experiments, they find the cheap talk script to significantly reduce the WTP values. Future research is to determine whether these attempts to reduce hypothetical bias will be valuable instruments in DCEs.

5.6.6 Follow-up questions

After the completion of the choice task, it is possible to include questions aimed at retrieving information on the respondents' reasons for answering the choice sets in the way that they did. These questions can relate to other validity aspects such as identification of response aberrations and of potential misunderstandings on the part of the respondent (Bennett & Blamey 2001):

- The examination of reasons for the observation of dominant preferences – for instance, are observed dominant preferences a result of 'true' strong (lexicographic) preferences or are they a symptom of decision-making heuristics?
- The examination of the cost attribute: existence of protest bias, payment vehicle bias, strategic bias and embedding (for a description of these terms, see chapter 6)
- The ability to understand the questionnaire
- The degree of difficulty in answering the questions – and reasons for this difficulty
- Further examination of preferences – i.e. the validity of the estimated preference function. This can be done, for instance, by including a ranking exercise in which the respondent is asked to rank the attributes by their importance on a Likert scale. The ranking exercise does not provide information about the respondents willingness to trade one attribute off for another (i.e. the results cannot be interpreted as cardinal utilities), but it does inform the

researcher as to whether or not the relative weighting of the attributes in the DCE is plausible

- Uncertainty question

Box 5.2.

Example of an uncertainty question
How sure are you about your decision regarding how much you would pay? Please circle one number between 1 and 10, where 1 indicates 'very unsure' and 10 indicates 'very sure'.

1	2	3	4	5	6	7	8	9	10
very unsure									very sure

Example of cheap talk
"In recent studies, several different groups of people wre asked to make choice like this. Although they were asked how much they thought they would pay, no one actually paid the money. These studies found that on average, people OVERSTATED their actual willingness to pay by as much as 1550 percent – quite a difference. In this question, please decide on how much you would pay exactly as if you were going to pay real money."

Source: (Samnaliev et al. 2003)

5.6.7 Socio-demographic questions

Finally, it is informative to include some questions on sociodemographic variables, such as age, sex, educational status, occupation, marital status, income, and so forth, as well as more specific questions related to the scenario presented in the task. The inclusion of sociodemographic questions makes it possible for the researcher to split the sample into subgroups and to test whether different groups of respondents answer differently, e.g. to test for heterogeneity in preferences. In practice this is done by examining interactions between personal characteristics and the desired attributes.

5.6.8 Focus group

Focus groups can have many different objectives and, conducted appropriately, they can contribute crucial information to the experiment. Use of a focus group allows the researcher to obtain insight into the respondents' way of thinking and to pilot the questionnaire on an appropriate group of individuals. Issues that can be successfully dealt with relate to all aspects of the study and include:

- General understanding of the questionnaire
- Complexity – degree of complexity versus cognitive burden
- Chosen attributes and their levels
- Cost attribute – reactions to the cost attribute (type of payment vehicle, appropriate level range etc.)
- Wording – such as framing effect (refer to chapter 6 for definition of framing effect)
- And so on……

Focus groups can be used in all the design stages of a DCE. They can be a cost efficient method of eliciting information – i.e. critical mistakes can be discovered early and a study that might otherwise turn out to be useless can be avoided. A problem with focus groups, however, can be their lack of representativeness for the survey population in question - individuals who agree to participate in a focus group may have different preferences, personal characteristics and cognitive abilities than the survey population as a whole.

Pre-testing refers to testing the questionnaire on a small sample of respondents to identify and correct potential problems before the main survey is undertaken. Bateman et al. (2002) argue that all DCEs should be pre-tested. In this context, focus groups can be seen as a good alternative to pilot testing when the latter is not an option - which is often the case due to money and time constraints.

5.6.9 Data collection method

When conducting a DCE it is important to consider the data collection procedure. Four main methods for collecting data exist (Bennett & Blamey 2001):

- Face-to-face interview
- Telephone interview
- Mailed questionnaires
- E-mail/internet
- Gathering in 'central facilities'
- Combination of the above

Face-to-face interviews are characterized by the interviewer and respondents sharing both time and space. Besides generating very high response rates, the advantage of this method is that the interviewer can lead the respondent through the hypothetical scenario and elaborate if the respondent does not understand the task. The NOAA panel has recommended that face-to-face interviews be used in contingent valuations as the benefits of this approach far exceed those of the other approaches (Arrow et al. 1993). There is a potential problem with personal interviews, however - namely interview bias, where the interviewer may influence the respondent's choices in an inadvisable manner. However, the major disadvantage in using face-to-face interviews is the cost, both in terms of money and time. The use of this data collection method has therefore been limited for DCEs in health care (e.g. Ryan & Gerard 2003).

In telephone interviews, the interviewer and the respondent share time, but not space. While telephone and personal interviews share many features, telephone interviews are a cheaper substitute. Due to the complexity of the DCE it is important that, besides the oral presentation of the questionnaire, the respondent is able to read the questions and see the scenarios (Curry 1997). It will be necessary, therefore, to mail the questionnaire to the respondent in advance of the telephone interview. This increases the cost - and thus reduces the advantages - of the method.

The mailed questionnaire approach is – for good reason – by far the most widespread data collection method for DCEs in many research areas, including health economics. Some of the advantages with this method are the relatively small cost compared to the amount of information gathered and the fact that respondents can chose to complete the questionnaire when it suits them. However, it also has some drawbacks. The method is prone to low response rates and thus sampling bias. A postal questionnaire limits the complexity of the choice task as the respondents must be able to answer the questions without help. Furthermore, the questionnaire needs to be written in simple language in order not to discriminate against individuals who are unused to completing forms and understanding written material. In studies where revealed preference data (mail survey) have been compared to DCE data (personal interview) (e.g. Haener et al. 2001;Samnaliev et al. 2003), the results suggest that surveys conducted via group sessions in which the interviewer is present obtain data of superior quality relative to the data from mail surveys (although the surveys were not conducted for this purpose). These authors argue that data of high quality are desirable as this will

increase the potential use of such data for benefit transfer – benefit transfer being one of the purposes of DCEs.

Computer-based interviewing is an approach that is gaining increasing recognition due to its ability to collect a lot of data relatively cheaply and to show complex designs in a simpler way, along with 3-D graphics that can often help to explain the hypothetical scenarios. A major disadvantage, however, is questionable representativeness of the final sample as there are still many people who do not feel confident using computers. Cairns & van der Pol (2004) used a web-based questionnaire in their examination of dominant preferences (in a study of time preferences for future health), in which the respondents were students from Aberdeen University. The use of computer-based data collection was crucial in this study as it allowed the choice sets (i.e. levels) to be varied according to the answers given previously, so as to reduce dominant responses. It is expected that web-based data collection will become more common in the future as access to and understanding of computers becomes more widespread. Gustafsson et al. (2000) suggested that computer-based interviewing will become more extensive, not only for financial reasons but also due to the demands arising from new ways of designing surveys.

5.6.10 The sample

The choice of sampling frame that is to be used to generate potential respondents (the survey sample) will depend upon the nature of the particular application. The sampling frame defines the universe of respondents from which a finite sample is drawn to whom the data collection instrument will be administered (Louviere et al. 2000). Examples of sampling frames are cancer patients, Danish citizens or pregnant women. The framing is dictated by the objective and perspective of the study. If the study intends to examine the use value of asthma medication, then the most appropriate frame would be asthma patients. If the aim of the study is to examine use as well as non-use and option value, however, then the appropriate frame would be the general public.

Based upon the sampling frame, the sampling strategy and sample size are determined. One sampling strategy is simple random sampling, in which all individuals from the sample frame have equal opportunity to be chosen as potential respondents; another sampling strategy might be dividing the frame into groups, each representing a portion of the population, depending upon characteristics such as sex, income, residential location etc. (Louviere et al. 2000). It is possible to determine the appropriate sample size by use of elementary statistics. This requires a priori knowledge about the choice probabilities, however, which is not straightforward as this information

is frequently not available[50]. The sample drawn depends on the need to split the sample into sub-samples. For instance, if a block design is used in the experimental design stage or/and validity tests are incorporated; the sample needs to be divided into sub-samples that represent the differences between the questionnaire versions. The size of the sample depends on the number of question given each respondent, the size of the population, and the statistical power that is required of the model derived. Bennett & Blamey (2001) state that the minimum size of a sub-sample should be in the order of 50 respondents, depending on the statistical power that is necessary for the estimation procedure. Furthermore, the sample size will be highly dependent on the expected response rate. To increase sample size (when using postal questionnaires), it might be appropriate to send out a reminder in the event of non-response.

5.7 Stage 5: Data analysis

Stage 5 includes organisation and typing of data, conduction of model estimation and policy analysis, and a critical discussion of the study and its results, i.e. a discussion of the validity and reliability of the study and comparison with other similar studies.

5.7.1 Data input

After the data are collected, they need to be organized and set into a computerized database. The choice of computer programme to use depends on personal preferences and experience, but both SPSS and Microsoft Excel ® are good candidates. Each choice set contains two forms of information:

1) The attribute levels of each alternative, and
2) Which of the given alternatives has been chosen

Related to 1), each attribute is handled as a variable containing different levels. Each attribute level has to be coded in order to estimate the importance of each attribute, i.e. the marginal values and trade of ratios. Data can be inputted in two ways. Either the attribute levels for each alternative can

[50] Ways to obtain such information include conducting a pilot study or reviewing results from revealed preference studies.

be inputted separately, or the difference in the attribute levels can be inputted[51]. The choice of strategy depends on how one prefers to handle the data and which software is to be applied. The outcome of the analysis will be the same, due to the fact that all probability models interpret the choice sets as the *difference* in the alternatives. For an example of coding of a choice set and data input, see Appendix II (the example serves only as illustration).

5.7.2 Data analysis

Data analysis involves two steps: first, the probability model is chosen and the data are run; secondly, the data are interpreted. The first step is discussed more fully in chapter 4 and is not discussed further here, except to note that there exist many computer programme that can handle probability models, including SAS, Stata, Limdep and Gauss. Appendix III gives a short example of data regression analysis.

When analysing the data it is important to take a critical approach as misinterpretation of the results can result in inefficient solutions and hence lead to undesirable policy implications. The next chapter considers some of the potential problems that relate to the DCE in general, as well as some of the ways in which DCE results can be validated.

[51] The last method only works for two alternatives per choice set.

6 Biases and validation

One of the advantages of using DCE and other stated preference methods (SP) is the ability to control the experiment. In comparison to market analysis (RP), the researcher is able to narrow the focus to the precise issue of interest. However, there is a potential danger in the DCE approach that the results will not be consistent with real-life settings due to possible differences between 'observed' and 'true' preferences. The hypothetical nature of DCE may therefore become a disadvantage and SP methods have been met with scepticism in many research areas. Within the health care field, critics of the DCE approach, such as Cookson (2003), argue that stated preference methods suffer from serious biases that render them unattractive to health care decision-makers and Wainwright (2003) states that:

> *"Of course, not all health-sector decisions have serious consequences for patients, for example, a service provider with a limited budget may have to choose whether to redecorate a patient waiting room or replace the furniture. In such circumstances CA [conjoint analysis - DCE] may be used without controversy to elicit consumer preferences […] However the stakes are raised significantly when a decision has profound consequences for the health and well-being of the patient, giving a strong impetus towards personal choice. Thus there appears to be an inverse relationship between the magnitude of a decision and the value of applying CA to it. […]Thus, rather than 'discovering' preferences that are already in existence, conjoint analysis [DCE] uses what is essentially a highly manipulative technique to construct a narrative or discourse about consumer preferences."* (Wainwright 2003, pp 378)

In comparison to methods that are based on actual behaviour, SP methods rely on the respondents' willingness *and* ability to reveal their true preferences. Thus the use of stated preference methods raises important questions - Will respondents answer honestly? Can they answer meaningfully? Do they understand the task? Can the results be generalized to other settings? Answering 'yes' to these questions is not at all straightforward, as examination of the performance of SP methods is difficult. One method that is gaining increasing interest involves the combination of RP and SP data. Such joint random utility models seem to be superior to either single RP or SP data modelling as they build on the strengths of both approaches (keeping in mind mutual disadvantages). See section ?? for further details.

As discussed throughout this paper, there are various sources of bias that can be associated with the DCE approach. These types of bias have been predominantly discussed within the psychology

literature – especially from authors such as D. Kahneman, A. Tversky and P. Slovic – with links to environmental economists such as J. Knetsch (e.g. Gilovich et al. 2002;Kahneman et al. 1982;Kahneman & Tversky 2000). Most of the literature on biases and validation in relation to stated preference methods is found in the environmental CVM literature, where the work by Mitchell and Carson (1989) can be characterized as the 'state of the art' in describing the experimental problems related to the CVM. Due to the similarity between DCE and CVM (and in particular dichotomous choice CVM), the methods share many of the same concerns related to biases. In the following sections a brief summary is given of the various biases that can occur and the different methods of testing for validity. The advantages and disadvantages of the CVM and DCE approaches are also discussed.

6.1 Biases

The validity and reliability of a DCE depends entirely on the magnitude of the associated biases. Thus it is extremely important when designing an experiment to try to minimize any biases that might exist. In economics, bias explains the difference between respondents' behaviour in a hypothetical market and in the real market, and the difference between theory and application. Gilovich et al. (2002) define biases as departures from the normative rational theory that serve as markers or signatures of the underlying heuristics. Each heuristic is thus associated with a set of biases. There has been a thorough discussion of biases in the CVM literature, which peaked in 1993 when the National Oceanic and Atmospheric Administration (NOAA) recommended a set of 'rules' when conducting CVM related to compensation for environmental disasters in USA. The panel summarized some of the potential problems that can arise in CVM studies (Arrow et al. 1993):

- CVM can produce results that appear to be inconsistent with assumptions of rational choice;
- Responses to CVM surveys sometimes seem implausibly large in view of the many programmes for which individuals might be asked to contribute and the existence of both public and private goods that might be substitutes for the resource(s) in question;
- Relatively few previous applications of the CVM method have reminded respondents forcefully of the budget constraints under which all must operate;
- It is difficult in CVM surveys to provide adequate information to respondents about the policy or programme for which values are being elicited and to be sure they have absorbed and accepted this information as the basis for their responses;

- In generating aggregate estimates using the CVM technique, it is sometimes difficult to determine the 'extent of the market'; and
- Respondents in CVM surveys may actually be expressing feelings about public spiritedness or the 'warm glow' of giving, rather than actual willingness to pay for the programme in question.

There is little consensus in the economic literature as how to classify and name the various biases; this is especially difficult as there are a large number of reported biases. The author has attempted to identify the most well-known, central biases that constitute a potential problem in the DCE – as well as in the CVM. Some examples of these biases are[52]:

1. *Incentives to misrepresent responses:*

- **Strategic bias**: Under- or over-estimation of WTP due to the pursuing of one's own interests (e.g. under-estimation: free-riding; over-estimation: securing of implementation of the intervention).
- **Justification bias**: Desire to rationalize the real-life choices made by the respondent. This especially relates to situations in which respondents expect researchers to possess information about how they really behave. Justification bias is thus likely to occur when SP and RP data are collected simultaneously.

2. *Scenario misspecification:*

- **Scope effect (scale effect):** When respondents do not distinguish between quantities of the good/attribute and hence value different quantities of a good/attribute equally.
- **Embedding effect (part-whole bias)**: When the value of the whole is different to the sum of the parts (Hanley et al. 1998). It is important to be aware that the terms 'embedding' and 'scope' are used interchangeably, although there according to some authors, e.g. Bateman et al. (2002), is an important distinction to be drawn between them. Where scope concerns a change in one argument of the multivariate utility function, embedding is concerned with the

[52] Unless otherwise indicated, the following is based on (Bateman et al. 2002;Bennett & Blamey 2001;Garrod & Willis 1999;Mitchell & Carson 1889).

change in at least two such arguments. One of the most prevalent arguments for the observation of scope/embedding effect is 'warm glow' where the desire may be to a considerable extent satiated by the very gesture of stating a willingness to contribute some reasonable amount, irrespective of quantity/characteristics of the good. For key literature exploring these issues see for example Beattie et al. (1998), Carthy et al. (1998) and Olsen et al. (2004)

- for empirical finding of scope effect in CVM in health care.
- **Status-quo bias (endowment effect)**: Preferences for the status-quo above other options (placing a higher value on goods once it is owned or one has experienced them). Salkeld et al. (2000) investigated this issue in a study of cancer testing programmes and found that consumers had a statistically significant preference for the existing services when all other factors remained constant. Moreover, Ryan & Ubach (2003) also found evidence for initial endowments of experience within health care.
- **Availability bias**: People's concerns and preferences are influenced by the publicity of the project/event in such a way that preferences appear irrational from a theoretical point of view. Jones Lee & Loomes (1994) reported that people's WTP to avoid a fatality on the London Underground was 69% greater than their WTP to avoid a road fatality.
- **Framing effect**: When respondents are unduly sensitive to the context or 'frame' in which a particular trade-off is offered. The main framing issue of concern relates to ensuring that respondents consider their budget constraints and substitute goods.
- **Ordering effects (sequencing)**: When the ordering of questions, attributes, etc. affects the attribute values. Within the health economics literature, Farrar & Ryan (1999) reported that the ordering of the attributes had no significant effect on the estimated utility weights; in contrast, Scott and Vick (1999) found evidence for an ordering effect.
- **Range bias**: The impact of the value estimates due to the choice of attribute levels (in CVM the choice of the range of WTP amounts). It has been recommended in the literature that the level chosen for attributes should be realistic, plausible and capable of being traded off (e.g. Ryan & Hughes 1997). Skjoldborg & Gyrd-Hansen (2003) found that a wider cost range led to an increase in WTP values.
- **Information bias**: Enough information has to be given to the respondents to enable them to make considered choices.

- **Payment vehicle bias**: The influence of the payment vehicle on the DCE choices made by the respondent. Payment vehicle bias has been reported in the health economics literature. Skjoldborg & Gyrd-Hansen (2003) compared two payment vehicles in their study of the Danish health care system and found that out-of-pocket payment had a higher negative impact on utility than the equivalent tax payment. It is important for the researcher to choose the appropriate payment vehicle so as to reduce potential bias and to best simulate real-life settings. San Miguel et al. (2000) commented that the use of cost to estimate WTP indirectly raises questions concerning the definition of the cost attribute in collectively funded health care systems. The application of an improper payment vehicle might ultimately induce protest bids, where respondents protest against the scenario.
- **Misspecification bias**: Inadequate communication so that the respondent does not perceive the information in the way intended by the researcher
- **Theoretical bias**: Violation of ex ante hypotheses and/or supporting economic theory. This phenomenon is discussed in more detail in the following section.

3. *Sampling and execution:*

- **Non-response bias**: The effect on the estimated values of total or part non-response of questions. In a study by Scott & Vick (1999), a response rate of less than 20% was reported; such a low response rate is likely to have implications for the representativeness of the sample, and hence the generalisability of the results. In order to examine the consequences of the low response rate, they compared some of the sample characteristics with those of the population as a whole and reported reasonable representativeness. High non-response rates lead to unreliable survey results (Arrow et al. 1993).
- **Sample selection bias**: Where the probability of an individual agreeing to be interviewed is related to the construct under investigation, leading to non-generalisability of results – for instance if people with higher cognitive abilities are more likely to participate in the survey.

6.2 Validity and reliability

Validity and reliability are crucial aspects when using techniques to elicit preferences, as preferences are not an observable phenomenon. This aspect is of utmost importance in stated

preference methods, as stated preferences attempt to simplify reality. It is important to understand the distinction between reliability and validity. Reliability refers to the degree of replicability of measurement over time and over different applications (i.e. generalisability), whereas validity refers to the extent to which a study measures what it intends to measure. Lack of validity can lead to systematic sources of error such as moral satisfaction, acquiescence, free-riding etc., whereas poor reliability can result in random error in the responses observed, i.e. reliability deals with whether a survey item measures something other than random noise (Bryan et al. 1998;Jorgensen et al. 2004)

The overall objective of validity is to measure the extent to which biases influence the estimates obtained. The validity of an experiment is thus the degree to which it measures the theoretical construct under investigation. The economic literature refers predominantly to three types of validity (described by the American Psychological Association in 1974): content validity, criterion validity and construct validity. It is important to bear in mind that not all three types of validity are equally important in every experiment. At the same time, there is no one type of validity that can act as a definite criterion for the validity of a stated preference method.

6.2.1 Content validity

Content validity (also termed face validity) refers to the extent to which an estimate takes account of all the issues deemed important for the experiment. This includes whether the DCE asks the correct questions in a clear, understandable, sensible and appropriate manner so that a valid estimate of the construct is obtained – e.g. the true WTP for an attribute. Content validity refers to all aspects of the experimental design, such as choice of attributes, attribute levels, ordering of attributes and framing. Content validity is difficult to assess, as it depends upon the intuitive judgement and experience of the person reviewing the study (subjective opinion from the researcher). The content validity of an experiment can be improved by ensuring thorough and careful selection of attributes, and by incorporating theory, use of experts and most importantly use of representative sample individuals in the study design. Focus groups and the conduction of pilot studies are key elements in improving the content validity of a study.

6.2.2 Criterion validity

Criterion validity (also termed external validity and sometimes subsumed under convergent validity) refers to the extent to which the measure of the construct is related to another measure

which may be regarded as a criterion; DCE estimates are typically compared with results from the actual market or from simulated market experiments. Since market prices are rarely available for public goods, criterion validity can generally only be tested against markets prices for private goods. However, this test may in many circumstances prove difficult to set up in real life. Carlsson and Martinsson (2001) argue that caution needs to be taken in performing tests of criterion validity that compare a private and a public (quasi-public) good, as there is a fundamental difference between private and public goods related to the incentives for truthful preference revelation. When a good is provided privately, the incentives for truthful preference revelation differ from the incentives associated with public goods in both an actual and a hypothetical context. For example, in a hypothetical setting, individuals may state what they believe is the market price or the fair price. Hence, caution should be given using private market goods to validate DCE for public (or quasi-public) goods, as the validity test will indicate an overestimation of the estimated WTP. When Carlsson and Martinsson (2001) tested for criterion validity, they found that values for wildlife protection estimated through DCE were insignificant from real payment estimates that were obtained in an experimental setting. The finding speaks in favour of incentive compatibility properties of DCEs (avoiding of hypothetical bias), however more research is needed within this field. See section 5.6.5 for a further discussion of hypothetical bias.

During the past decade the interest in combining RP and SP data in a RUT-based modelling framework has increased considerably – with DCE being especially appropriate for joint modelling. While RP data can help to illuminate preferences within existing or recent past market/technology structures, SP data can provide information about how preferences are likely to respond to shifts in technology frontiers. This suggests that RP and SP data sources have complementary strengths and weaknesses which can be exploited to enhance the understanding of preference processes – hence enhancing validity (Hensher et al. 1999). For additional information, refer to section 4.9.

6.2.3 Construct validity

Construct validity refers to the extent to which a particular measure relates to other measures that are consistent with the theoretically-derived hypotheses that relate to the concept being measured. Construct validity is concerned with two aspects. First, whether the measure is correlated with other measures of the same theoretical constructs – named *convergent validity*. Second, whether the measure is related to measures of other constructs in a manner predicted by theory (such as economic theory) or sometimes, more generally, a priori expectations – named *theoretical*

validity. It is important to be aware of the difference between criterion validity and convergent validity. Criterion validity is a comparison with an external criterion, which is closer to the theoretical construct of the investigation, whereas convergent validity is a comparison with other types of elicitation methods that measure the same construct.

Convergent validity

Convergent validity (also termed external validity and sometimes countering criterion validity) refers to the correspondence or convergence between measures obtained by different preference-based methods. It would be appropriate, therefore, to compare DCE estimates with estimates obtained from another stated preference method (such as CVM), other choice techniques (ranking/rating), multiple DCE studies, or from a RP method such as hedonic pricing or travel cost. To successfully compare such results, the measured value must be the same in each case. For instance, when comparing the DCE with a revealed preference method, the clarification of non-use value and option value is essential as the RP approach does not include such values. A further possibility exits within health care, where DCE estimates can be tested with respect to other evaluation methods such as the standard gamble, time-trade-off and visual analogue scale approaches[53] (Drummond et al. 1997). In convergent validity testing, no measure is superior to the other in terms of being a closer approximation to the value of the underlying construct. Two experiments that deliver the same estimates are not necessarily both valid – they might just be equally invalid. However, a large and unexpected difference between the estimates from two experiments would suggest that at least one of the measures is invalid.

One of the first attempt to compare dichotomous choice CVM and DCE estimates was a study of recreational moose hunting ((Boxall et al. 1996). These authors found that the welfare estimate differed between the two approaches; the reason for the difference, however, was most likely that respondents did not consider substitution possibilities in the case of CVM - the difference was thus not due to behavioural differences but to framing effects. The study by Cameron et al. (2002) provided one of the most extensive comparisons of different SP methods. Five methods were compared with DCE, including: actual choice, two versions of dichotomous choice CVM, open-ended CVM, payment card and multiple bounded[54]. Instead of comparing WTP point estimates, they compared the underlying preference functions, taking scale differences into consideration. They found that four of the six methods had significantly identical utility functions. The two

[53] Methods for evaluating of health status – permitting the calculation of QALYs
[54] Respondents were asked to indicate how likely they would be to pay on a scale from 1 to 5

methods that were least consistent with the others were open-ended CVM and payment card (which is in line with general research), suggesting that the decision-making process (the use of heuristics) differs for these two approaches. Boyle et al. (2001) reported a similar result when comparing the three choice techniques (DCE, ranking and rating). Their results indicated that convergent validity was not established for these methods. On these grounds the authors recommended that ratings not be recoded to ranking or DCE, and that DCE should be the method of choice, as this avoids the concerns of cardinality and provides the most conservative welfare estimate.

According to Ryan & Gerard (2003), four health care studies have addressed convergent validity by comparing the elicited value estimates with QALY estimates based on standard gamble, time trade-off and visual analogue scale approaches. In a study of in vitro fertilisation Ryan (2004a) compared WTP estimates obtained from dichotomous choice CVM and DCE and found that the derived WTP estimates did not significantly differ from each other. Mean WTP for DCE lay within the 95% confidence intervals for WTP for dichotomous choice - and vice versa. This suggests high validity of comparable value estimates of DCE and dichotomous choice CVM, but does not validate the overall use of SP methods, however.

Theoretical validity

Theoretical validity (also termed internal validity) refers to extent to which the findings conform to the theoretical foundation of the experiment and to *a priori* expectations. In this case the focus is on the determinants of the estimates rather than, as was the case with convergent validity, the fit between two separate but equal measures of the same construct. Violation of the economic theory is of utmost importance if the value estimates are to be used in an economic context, such as in cost-benefit analysis. Severe violations of theory greatly restrict the interpretation of the results and leave the researcher with little to work on. DCE is particularly appropriate for theoretical validity testing as it normally consists of multiple choices (Hanley et al. 1998). Theoretical validity testing has been used widely in health care (e.g. Ryan & San Miguel 2003). Tests related to the axioms of economic theory are especially well reported, including ex ante hypothesis of negative utility for price and diminishing marginal utility of income, and approaches such as consistency tests and tests of dominant preferences (whether individuals employ compensatory decision-making).

Consistency test

The consistency test (also referred to as internal consistency test) relates to whether the respondents understand the concept of the DCE and hence the extent to which they act rationally (according to economic theory) when expressing their preferences. The test is performed by including a scenario in which one alternative is unquestionably dominated by the other (i.e. alternative *a* should be preferred to alternative *b*). A superior but more data-demanding method is to test for rationality by use of transitivity tests (i.e. if alternative *a* is preferred to alternative *b*, and alternative *b* is preferred to alternative *c*, then alternative *a* should be preferred to alternative *c*). There has been much debate on the issue of consistency and how to deal with inconsistent answers. The discussion has its base in the theory related to the rationality of consumers in general and how they make their trade-offs (e.g. McFadden 1999). On one side is the notion that *a preference is a preference no matter what* - and no one can be a judge of whether an answer is inconsistent or not. On the other side, the psychology literature argues that inconsistency is due to cognitive inabilities and the use of heuristics in the decision-making process (Gilovich et al. 2002;Kahneman et al. 1982). McIntosh & Ryan (2002) argue that the important issue when discussing internal consistency is whether irrational *responses* represent irrational *respondents*. A low number of inconsistent answers may suggest that these 'intransitive' choice sets are due to a mistake on the part of the respondent and may not be based on irrationality. Consistency and transitivity tests have been conducted in health care – and with varying results (e.g. Bryan et al. 1998; Farrar et al. 2000; McIntosh & Ryan 2002; Ratcliffe & Buxton 1999; Ryan 1999; Ryan & Hughes 1997; San Miguel et al. 2000; Scott & Vick 1999; van der Pol & Cairns 2001).

Testing for dominant preferences

Special attention has been given to the appearance of dominance of specific attributes as this might influence the results. Dominance refers to a situation in which a respondent persistently chooses the alternative containing the best level of a particular attribute. If such a behaviour pattern is identified, the respondent is said to possess dominant preferences for that particular attribute[55], e.g. in the case of a respondent who consistently chooses the cheapest treatment. Such a respondent does not make trade-offs (the respondent does not adopt compensatory decision-making) – at least not for the given level range. By definition, dominance places no restrictions on the ordering or

[55] It is possible to perform a pseudo t-test in which a particular attribute parameter estimated in a standard model (including all variables) is compared with the same attribute parameter estimated in a model including only this variable. If the t-test is accepted (H: $\beta=\beta$) it is an indication of dominance

extent of trading for other attributes. There are many possible reasons why dominant preferences are observed in DCEs. According to Scott (2002), dominant preferences might be due to:

- Task complexity: the respondents do not have the cognitive ability to process the information due to conflicts between values, such as trade-offs.
- Too high a number of choice sets and/or attributes: the respondents cannot cope with the task and thus uses a rule of thumb as a decision rule. Both this and task complexity result in employment of simple decision-making heuristics.
- The design: if the attribute levels are constructed 'wrongly', the respondents will not trade between levels simply because the interval between the levels is too great.
- Strong preferences: the observation of dominance may reflect very strong preferences and a belief that a specific attribute is the important one; respondents were thus not prepared to make trade-offs. This implies a *lexicographic ordering* and could indicate a 'right-based' view of choice and an ethical belief that individuals should be provided with a specific characteristic. In this case, indifference curves cannot be formulated as trading does not take place and MRS is thus meaningless (Deaton & Muellbauer 1989).

Various health care DCE studies have examined the influence of dominant preferences. Bryan et al. (1998) found a high proportion of lexicographic preferences for one of the attributes, suggesting the use of simple decision-making heuristics. Cairns et al. (2002) - using the same data – examined whether dominance indicated employment of simple decision-making heuristics (i.e. fast and frugal heuristics) or whether dominance was due to the attribute level ranges offered. A follow-up method was applied in which individuals were presented with trade-offs that varied according to previous answers. Only one individual was found not to make trade-offs between attributes, which was highly suggestive of dominance being a question of choosing the right design, i.e. the appropriate range in attribute levels. The extent to which dominance is observed varies considerably in DCE studies, and 50% of response dominance is not unusual. Currently, there is no consensus of how to interpret and handle these responses. Future studies are needed to examine the reasons for dominance behaviour.

6.3 CVM versus DCE

The distinction between CVM (in particular dichotomous choice CVM) and DCE is not always clear as the two approaches are very similar. Nevertheless, it is the small differences that give rise to discussion as to which method is preferable in a given situation and which is more robust with respect to bias. Some biases that can be quite problematic in CVM do not seem to create big problems in DCE –examples include the issues of protest bids and starting point bias. It has also been argued that DCE more closely resembles everyday life which are likely to increase incentive compatibility (Lusk & Schroeder 2004). Hanley et al. (1998) and Bennett & Blamey (2001) argue that DCE possesses several important advantages, such as:

- The ability to estimate attribute values and hence marginal effects
- Better suited for benefit transfer[56]
- Avoiding of 'yea' saying (compliance bias)
- Less embedding/scope effect: in general there is a significant body of evidence to suggest that CVM estimates exhibit great insensitivity to scope, whereas DCE has a greater sensitivity
- Allowance for validity tests such as consistency tests

Rolfe et al. (2002) consider DCE to have significant framing advantages over CVM as it allows respondents to choose from a large pool of potentially substitutable goods (respondents are asked to choose between different alternatives) and is a more realistic way for respondents to trade-off opportunity costs. In comparison, CVM assumes that respondents have considered the alternative and substitute goods, although the researcher has no direct evidence that they have. The ability in DCE to disguise the good in question within a pool of potential trade-offs is an important way of minimizing information transfer and other potential biases, and of modelling realistic choices. This may be especially important when the good in question is not very familiar to respondents. The DCE also allows a more rigorous testing of framing effects than does the CVM, as the analyst can test whether framing differences cause variation in parameters. For example, Rolfe et al. (2002) found evidence for framing effects in relation to substitutes. These authors argued that even though DCEs possess real advantages over CVM in relation to framing effects, these framing effects are

[56] In health care, (Johnson et al. 2000) argue that the use of generic health attributes facilitates the transfer of WTP estimates for CBA of a variety of potential health interventions.

more widespread in DCE than may be commonly thought. Moreover the pool of substitutes and choice options offered might influence the extent of framing.

An advantage of DCE is that biases can be minimized, as the good of interest can be 'hidden' within the pool of available goods. The cost attribute is thus one of several attributes that describe the good in question. By de-emphasising the importance of this attribute relative to its central role in CVM biases may be reduced. and its importance is diminished, in comparison to its central role in a CVM.

Although DCE is in many ways a superior method to CVM one should not forget that it also has several drawbacks. According to Carlsson & Martinsson (2001), and in line with the methodological issues discussed in this paper, it can be argued that:

- DCE is much more demanding for respondents to answer
- Preferences may be unstable throughout the experiment
- The incentive properties might be unclear
- The designing of experiments is a difficult task (the statistical/experimental design is often complex; selection of appropriate attributes and levels can be difficult)

Bateman et al. (2002) note that it is more difficult to apply DCE than CVM to derive values for a sequence of elements implemented by policy or projects (i.e. valuation of a package of goods). While DCE is not as sensitive to scope effect as CVM, it is standard in DCE to assume a linear utility function in which 'the value of the whole is equal to the sum of the parts'. This raises two concerns: firstly, regarding the influence of attributes not included in the experiment (and thus not contributing to utility) and, second, whether the linear additive assumption is appropriate. Foster & Mourato (2003) made a systematic comparison of dichotomous choice DCE and CVM with emphasis on embedding/scope effect and the linearity assumption. The study was conducted to examine the social value of services provided by the charitable sector in the U.K – such as that related to housing and homelessness, social services and health research. These authors also questioned the assumption of a linear additive utility function. They found that the value of the good as a whole was larger in the DCE than in the CVM[57] and thus suggested that respondents perceive CVM and DCE differently and that care must be taken to choose the most appropriate

[57] However, whether CVM potentially understates or DCE potentially overstates true WTP remains an open question

methodology. They recommended restricting the use of DCE to the evaluation of a single isolated change or of a policy that is part of a larger set of changes/policies. Carlsson & Martinsson (2001) argued that the choice of SP method depends firstly on the context of the experiment and secondly on the trade-off between the advantages and disadvantages of the methods. Among other things, it might be difficult to describe the attributes and/or include all possible attributes. For example, there are many complex attributes associated with the valuation of a natural habitat – including biodiversity, recreation, environmental protection, employment, food etc. The wrongly exclusion of important attributes would be critical to the analysis and hence the CVM approach might be preferable in such cases –although the opportunity to estimate marginal effects would be lost. As accurately expressed by Boyle et al. (2001):

> *"In fact they [SP methods] are so closely related that it would be surprising if any one approach proves to be a panacea for all the problems that have been asserted to apply to stated preference methods*" (Boyle et al. 2001), pp 441)

7 Conclusion

> *"If [CV] and [DCE] are to be extended to the health policy arena, it is important that methodology continues to be improved. Given the policy success of such methods within environmental economics, it is important that we learn from, and take note of, lessons that con be learnt from this sub-discipline."* (Hanley et al. 2003, pp 4).

In this paper, the discrete choice experiment has been placed into the context of preference-based economic evaluation and issues relating to theory, design, modelling and biases have been reviewed, with emphasis on its application within health care. The DCE is increasingly being used within health care and appears to have some advantages over the CVM as it may minimize some sources of bias and provides the researcher with an opportunity to examine the individual impact of the characteristics that make up the good or service in question - hence increasing the amount of information obtained by the researcher. DCE also has an advantage with respect to validation of economic theory - not only the standard neoclassical economic theory but also others such as principal-agent theory and transaction cost theory.

A difficulty in relation to validation is that many of the methodological issues that are raised in relation to economic theory appear to be study-specific, thus making it difficult to generalize the findings. Although examples exist – especially in the psychology literature – of economic theory not reflecting real-life situations, this does not justify the rejection of economic theory. It is still important to have a framework that allows accurate predictions of consumer behaviour – the absence of a framework would mean the inability to make ex ante hypotheses. Departures from economic theory are not desirable and undermine the power of the theory, but they can be accepted as long as the chances of the theory being appropriate are greater than it being inappropriate.

Despite increasing use of DCE in the health field, health economists still have much to learn about the features of DCE and related methods. It is still very much a technique under development in this field. Other fields – such as environmental and transportation economics – have a longer history of using stated preference methods and discrete choice modelling and are considerably further ahead in their research of these techniques. It is clear that experimental design plays a vital role in the performance of a DCE. Effort has thus been spent on determining optimal design and examining the relationship between design and issues such as complexity of the choice task and cognitive demands. If deviation from economic rationality is due to the cognitive burden placed on respondents, it is very important in the future to ensure familiarity with the scenario content and to

minimize the task complexity. Further research is also needed to examine the relationships between modelling, experimental design and individual judgement and decision-making. The behavioural processes that are associated with the DCE are now better understood, but much more needs to be learned about the method if it is to be a useful tool for future evaluations.

Furthermore, it has been the author's intention to present some of the most important and newly recognized ideas and concerns in the application of the DCE. It is important to be aware that the DCE is a 'moving target' and any attempt at defining the golden standard methodology in relation to DCE would be premature.

References

Aadland, D. & Caplan, A. J. Cheap Talk Revisited: New Evidence for CVM. Utah State University. Working Paper: eri0220 . 2003.

Adamowicz, W., Louviere, J., & Williams, M. 1994, "Combining Revealed and Stated Preference Methods for Valuing Environmental Amenities", *Journal of Environmental Economics and Management*, vol. 26, no. 3, pp. 271-292.

Adamowicz, W., Louviere, J. J., & Swait, J. 1998, *Introduction to Attribute-Based Stated Choice Models*, US Department of Commerce, NOAA.

Anderson, N. H. 1962, "Application of An Additive Model to Impression Formation", *Science*, vol. 138, no. 3542, p. 817-&.

Anderson, S., Palma, A., & Thisse, J. 1991, *Discrete Choice Theory of Product Differentiation,* The MIT Press, Cambridge, England.

Arrow, K., Solow, R., Portney, P., Leamer, E., Radner, R., & Schuman, H. 1993, *Report of the NOAA Panel on Contingent Valuation*, NOAA, U.S. Department of Commerce, USA.

Bateman, I. J., Carson, R. T., Day, B., Hanemann, M., Hanley, N., Hett, T., Jones-Lee, M., Loomes, G., Mourato, S., Özdemiroglu, E., Pearce, D., Sugden, J., & Swanson, J. 2002, *Economic Evaluation with Stated Preference Techniques, A Maunal*, 1 edn, Edward Elgar Publishing Limited, Cheltenham.

Beattie, J., Covey, J., Dolan, P., Hopkins, L., Jones-Lee, M., Loomes, G., Pidgeon.N., Robinson, A., & Spenser, A. 1998, "On the Contingent Valuation of Safety and the Safety of Contingent Valuation: Part 1-Caveat Investigator", *Journal of Risk and Uncertainty*, vol. 17, no. 1, pp. 5-26.

Bech, M. 2003, "Politicians' and hospital managers' trade-offs in the choice of reimbursement scheme: a discrete choice experiment", *Health Policy*, vol. 66, no. 3, pp. 261-275.

Ben-Akiva, M. & Lerman, S. 1985, *Discrete Choice Analysis: Theory and application to travel demand*, 2 edn, The MIT Press, Cambridge, Massachusetts.

Ben-Akiva, M., McFadden, D., Abe, M., Böckenholt, U., Bolduc, D., Gopinath, D., Morikawa, T., Ramaswamy, V., Rao, V., Revelt, D., & Steinberg, D. 1997, "Modeling Methods for Discrete Choice Analysis", *Marketing Letters*.

Ben-Akiva, M., McFadden, D., Gärling, T., Gopinath, D., Walker, J., Bolduc, D., Börsch-Supan, A., Delquié, P., Larichev, O., Morikawa, T., & Polydoropoulou, A. 1999, "Extended Framework for Modeling Choice Behavior", *Marketing Letters*, vol. 10, no. 3, pp. 187-203.

Ben-Akiva, M. & Morikawa, T. 1990, "Estimation of Switching Models from Revealed Preferences and Stated Intentions", *Transportation Research Part A-Policy and Practice*, vol. 24, no. 6, pp. 485-495.

Bennett, J. & Blamey, R. K. 2001, *The choice Modelling approach to Environmental Valuation* Edward Elgar Publishing Limited, UK.

Bhat, C. R. & Castelar, S. 2002, "A unified mixed logit framework for modeling revealed and stated preferences: formulation and application to congestion pricing analysis in the San Francisco Bay area", *Transportation Research Part B: Methodological*, vol. 36, no. 7, pp. 593-616.

Blamey, R. K., Bennett, J. W., Louviere, J. J., Morrison, M. D., & Rolfe, J. 2000, "A test of policy labels in environmental choice modelling studies", *Ecological Economics*, vol. 32, no. 2, pp. 269-286.

Blamey, R. K., Bennett, J. W., Louviere, J. J., Morrison, M. D., & Rolfe, J. C. 2002, "Attribute causality in environmental choice modelling", *Environmental & Resource Economics*, vol. 23, no. 2, pp. 167-186.

Boardman, A. E., Greenberg, D. H., Vining, A. R., & Weisbrod, B. A. 2001, *Cost-Benefit Analysis*, 2 edn, Printice Hall, Upper Saddle River, New Jersey.

Boardway, R. W. & Bruce, N. 1984, *Welfare Economics* Basil Blackell Publisher Limited, Oxford.

Boxall, P. C. & Adamowicz, W. L. 2002, "Understanding heterogeneous preferences in random utility models: A latent class approach", *Environmental & Resource Economics*, vol. 23, no. 4, pp. 421-446.

Boxall, P. C., Adamowicz, W. L., Swait, J., Williams, M., & Louviere, J. 1996, "A comparison of stated preference methods for environmental valuation", *Ecological Economics*, vol. 18, no. 3, pp. 243-253.

Boxall, P. C., Englin, J., & Adamowicz, W. L. 2003, "Valuing aboriginal artifacts: a combined revealed-stated preference approach", *Journal of Environmental Economics and Management*, vol. 45, no. 2, pp. 213-230.

Boyle, K. J., Holmes, T. P., Teisl, M. F., & Roe, B. 2001, "A comparison of conjoint analysis response formats", *American Journal of Agricultural Economics*, vol. 83, no. 2, pp. 441-454.

Brainard, J., Bateman, I., & Lovett, A. 2001, "Modelling demand for recreation in English woodlands", *Forestry*, vol. 74, no. 5, pp. 423-438.

Bryan, S., Buxton, M., Sheldon, R., & Grant, A. 1998, "Magnetic resonance imaging for the investigation of knee injuries: An investigation of preferences", *Health Economics*, vol. 7, no. 7, pp. 595-603.

Cairns, J. & van der Pol, M. 1997, "saving future lives. A comparison of three discoutning models", *Health Economics*, vol. 6, no. 4, pp. 341-350.

Cairns, J. & van der Pol, M. 2004, "Repeated follow-up as a method for reducing non-trading behaviour in discrete choice experiments", *Social Science & Medicine*, vol. 58, no. 11, pp. 2211-2218.

Cairns, J., van der Pol, M., & Lloyd, A. 2002, "Decision making heuristics and the elicitation of preferences: Being fast and frugal about the future", *Health Economics*, vol. 11, no. 7, pp. 655-658.

Cameron, T. A., Poe, G. L., Ethier, R. G., & Schulze, W. D. 2002, "Alternative non-market value-elicitation methods: Are the underlying preferences the same?", *Journal of Environmental Economics and Management*, vol. 44, no. 3, pp. 391-425.

Carlsson, F., Frykblom, P., & Lancaster, K. J. Using Cheap-Talk as a Test of Validity in Choice Experiments. 2004. Department of Economics, Gothenburg University. Working Papers in Economics no. 128.

Carlsson, F., Frykblom, P., & Liljenstolpe, C. 2003, "Valuing wetland attributes: an application of choice experiments", *Ecological Economics*, vol. 47, no. 1, pp. 95-103.

Carlsson, F. & Martinsson, P. 2001, "Do hypothetical and actual marginal willingness to pay differ in choice experiments? Application to the valuation of the environment", *Journal of Environmental Economics and Management*, vol. 41, no. 2, pp. 179-192.

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

Carson, R. T. 1999, *Contingent Valuation: A User's Guide; Discussion Paper* University if Califonia, San Diego, Department of Economcis.

Carthy, T., Chilton, S., Covey, J., Hopkins, L., Jones-Lee, M., Loomes, G., Pidgeon.N., & Spenser, A. 1998, "On the Contingent Valuation of Safety and the Safety of Contingent Valuation: Part 2 - The CV/SG "Chained" Approach", *Journal of Risk and Uncertainty*, vol. 17, no. 3, pp. 187-214.

Champ, P. A. & Bishop, R. C. 2001, "Donation payment mechanisms and contingent valuation: An empirical study of hypothetical bias", *Environmental & Resource Economics*, vol. 19, no. 4, pp. 383-402.

Clarke, P. M. 1998, "Cost-benefit analysis and mammographic screening: a travel cost approach", *Journal of Health Economics*, vol. 17, no. 6, pp. 767-787.

Clarke, P. M. 2002, "Testing the convergent validity of the contingent valuation and travel cost methods in valuing the benefits of health care", *Health Economics*, vol. 11, no. 2, pp. 117-127.

Cookson, R. 2003, "Willingness to pay methods in health care: a sceptical view", *Health Economics*, vol. 12, no. 11, pp. 891-894.

Cummings, R. G. & Taylor, L. O. 1999, "Unbiased value estimates for environmental goods: A cheap talk design for the contingent valuation method", *American Economic Review*, vol. 89, no. 3, pp. 649-665.

Curry, J. 1997, "After the Basic: Keeping Key issues in mind makes conjoint analysis easier to apply", *Marketing Research*, vol. 9, pp. 6-11.

Deaton, A. & Muellbauer, J. 1989, *Economics and consumer behaviour* Cambridge Univeristy Press, USA.

Dellaert, B., Brazell, J. D., & Louviere, J. J. 1999, "The Effect of Attribute Variation on Consumer Choice Consistency", *Marketing Letters*, vol. 10, no. 2, pp. 139-147.

Delucchi, M. A., Murphy, J. J., & McCubbin, D. R. 2002, "The health and visibility cost of air pollution: a comparison of estimation methods", *Journal of Environmental Management*, vol. 64, no. 2, pp. 139-152.

Diener, A., O'Brien, B., & Gafni, A. 1998, "Health care contingent valuation studies: A review and classification of the literature", *Health Economics*, vol. 7, no. 4, pp. 313-326.

Dolan, P. & Edlin, R. 2002, "Is it really possible to build a bridge between cost-benefit analysis and cost-effectiveness analysis?", *Journal of Health Economics*, vol. 21, no. 5, pp. 827-843.

Drummond, M. F., O'Brian, B., Stoddart, G. L., & Torrance, G. W. 1997, *Methods for the Economic Evalueation of Health Care Programmes*, second edn, Oxford Medical Publications, Oxford University Press, U.K.

Farrar, S., Ryan, M., Ross, D., & Ludbrook, A. 2000, "Using discrete choice modelling in priority setting: an application to clinical service developments", *Social Science & Medicine*, vol. 50, no. 1, pp. 63-75.

Foster, V. & Mourato, S. 2003, "Elicitation format and sensitivity to scope - Do contingent valuation and choice experiments give the same results?", *Environmental & Resource Economics*, vol. 24, no. 2, pp. 141-160.

Freeman, A. M. 1999, *The Measurement of Enviromental and Resource Values*, 3 edn, Resources for the Future, Washington, DC.

Garrod, G. & Willis, K. 1999, *Economic Valuation of the Environment, Methods and Case Studies* Edward Elgar Publishing Limited, Cheltenham UK.

Garrod, G. D. & Willis, K. G. 1992, "Valuing Goods Characteristics - An Application of the Hedonic Price Method to Environmental Attributes", *Journal of Environmental Management*, vol. 34, no. 1, pp. 59-76.

Gilovich, T., Griffin, D., & Kahneman, D. 2002, *Heuristics and Biases* Cambrigde University Press, UK.

Green, D., Jacowitz, K. E., Kahneman, D., & McFadden, D. 1998a, "Referendum contingent valuation, anchoring, and willingness to pay for public goods", *Resource and Energy Economics*, vol. 20, no. 2, pp. 85-116.

Green, D., Jacowitz, K. E., Kahneman, D., & McFadden, D. 1998b, "Referendum contingent valuation, anchoring, and willingness to pay for public goods", *Resources and Energy Economics*, vol. 20, no. 2, pp. 85-116.

Green, P. E., Carmone, F. J., & Wind, Y. 1972, "Subjective Evaluation Models and Conjoint Measurement", *Behavioral Science*, vol. 17, no. 3, pp. 288-299.

Green, P. E. & Rao, V. R. 1971, "Conjoint Measurement for Quantifying Judgmental Data 1", *Journal of Marketing Research*, vol. 8, no. 3, pp. 355-363.

Green, P. E. & Srinivasan, V. 1978, "Conjoint Analysis in Consumer Research - Issues and Outlook 2", *Journal of consumer research*, vol. 5, no. 2, pp. 103-123.

Green, P. E. & Srinivasan, V. 1990, "Conjoint Analysis in Marketing: New Developments With Implications for Research and Practice", *Journal of Marketing*, vol. 54, no. 4, pp. 3-19.

Greene, W. H. 2003, *Econometric Analysis*, 5 edn, Person Education, Inc., New Jersey, USA.

Gustafsson, A., Herrmann, A., & Huber, F. 2000, *conjoint Measurement - Methods and Application* Springer-Verlag, Berlin, Germany.

Gyrd-Hansen, D. 2003, "Willingness to pay for a QALY", *Health Economics*, vol. 12, no. 12, pp. 1049-1060.

Gyrd-Hansen, D. & Slothuus, U. 2002, "The citizen's preferences for financing public health care: a Danish survey", *Int.J.Health Care Finance Econ.*, vol. 2, no. 1, pp. 25-36.

Haan, P. Discrete Choice Labor Supply: Conditional logit vs. Random Coefficient Models. Discussion Paper 394. DIW Berlin.German Institute for Economic Research. 2004.

Haener, M. K., Boxall, P. C., & Adamowicz, W. L. 2001a, "Modeling recreation site choice: Do hypothetical choices reflect actual behavior?", *American Journal of Agricultural Economics*, vol. 83, no. 3, pp. 629-642.

Hall, J., Kenny, P., King, M., Louviere, J., Viney, R., & Yeoh, A. 2002, "Using stated preference discrete choice modelling to evaluate the introduction of varicella vaccination", *Health Economics*, vol. 11, no. 5, pp. 457-465.

Hall, J., Viney, R., Haas, M., & Louviere, J. 2004, "Using stated preference discrete choice modeling to evaluate health care programs", *Journal of Business Research*, vol. 57, no. 9, pp. 1026-1032.

Hanemann, W. M. 1994, "Valuing the Environment Through Contingent Valuation", *Journal of Economic Perspectives*, vol. 8, no. 4, pp. 19-43.

Hanemann, W. M. 1984, "Welfare Evaluations in Contingent Valuation Experiments with Discrete Responses", *American Journal of Agricultural Economics*, vol. 66, no. 3, pp. 332-341.

Hanemann, W. M. 1991, "Willingness to Pay and Willingness to Accept - How Much Can They Differ", *American Economic Review*, vol. 81, no. 3, pp. 635-647.

Hanemann, W. M. 1995, "Contingent Valuation and Economics," in *Environmental Economics - New Perspectives*, K. Willis & J. Corkindale, eds., CAB International, pp. 79-117.

Hanley, N., Ryan, M., & Wright, R. 2003, "Estimating the monetary value of health care: lessons from environmental economics", *Health Economics*, vol. 12, no. 1, pp. 3-16.

Hanley, N. & Spach, C. L. 1993, *Cost-Benefit Analysis and the Environment* Edward Elgar Publishing Company, England.

Hanley, N., Wright, R. E., & Adamowicz, V. 1998, "Using choice experiments to value the environment - Design issues, current experience and future prospects", *Environmental & Resource Economics*, vol. 11, no. 3-4, pp. 413-428.

Hanley, N., Wright, R. E., & Koop, G. 2002, "Modelling recreation demand using choice experiments: Climbing in Scotland", *Environmental & Resource Economics*, vol. 22, no. 3, pp. 449-466.

Hensher, D. & Greene, W. H. 2002, *The Mixed Logit Model: The State of Practice. Working Paper*, Institute of Transport Studies, University of Sydney.

Hensher, D., Louviere, J., & Swait, J. 1999, "Combining sources of preference data", *Journal of Econometrics*, vol. 89, no. 1-2, pp. 197-221.

Hesseln, H., Loomis, J. B., Gonzalez-Caban, A., & Alexander, S. 2003, "Wildfire effects on hiking and biking demand in New Mexico: a travel cost study", *Journal of Environmental Management*, vol. 69, no. 4, pp. 359-368.

Hidona, N. 2002, *The Economic Valuation of the Environment and Public Policy - A Hedonic Approach* Edward Elgar Publishing Limited, Cheltenham, UK.

Huber, J. & Zwerina, K. 1996, "The importance of utility balance in efficient choice designs", *Journal of Marketing Research*, vol. 33, no. 3, pp. 307-317.

Hundley, V., Ryan, M., & Graham, W. 2001, "Assessing women's preferences for intrapartum care", *Birth-Issues in Perinatal Care*, vol. 28, no. 4, pp. 254-263.

Hutton, J. & Maynard, A. 2000, "A nice challenge for health economics", *Health Economics*, vol. 9, no. 2, pp. 89-93.

Iraguen, P. & Dios Ortuzar, J. 2004, "Willingness-to-pay for reducing fatal accident risk in urban areas: an Internet-based Web page stated preference survey", *Accident Analysis and Prevention*, vol. 36, no. 4, pp. 513-524.

Jan, S., Mooney, G., Ryan, M., Bruggemann, K., & Alexander, K. 2000, "The use of conjoint analysis to elicit community preferences in public health research: a case study of hospital services

in South Australia", *Australian and New Zealand Journal of Public Health*, vol. 24, no. 1, pp. 64-70.

Johansson, P. O. 1987, *The economic theory and measurement of environmental benefits* Cambridge University Press, Cambridge.

Johansson, P. O. 1993, *Cost-benefit analysis of environmental change* Cambridge University Press, Cambridge.

Johansson.P. 1991, *An introduction to modern welfare economics* Cambridge University Press, Cambridge.

Johansson.P.O. 1991, *An introduction to modern welfare economics* Cambridge University Press, Cambridge.

Johnson, F. R., Banzhaf, M. R., & Desvousges, W. H. 2000, "Willingness to pay for improved respiratory and cardiovascular health: A multiple-format, stated-preference approach", *Health Economics*, vol. 9, no. 4, pp. 295-317.

Johnson, F. R., Desvousges, W. H., Ruby, M. C., Stieb, D., & De Civita, P. 1998, "Eliciting stated health preferences: An application to willingness to pay for longevity", *Medical Decision Making*, vol. 18, no. 2, p. S57-S67.

Jones Lee, M. & Loomes, G. 1994, "Towards A Willingness-To-Pay Based Value of Underground Safety", *Journal of Transport Economics and Policy*, vol. 28, no. 1, pp. 83-98.

Jorgensen, B. S., Syme, G. J., Smith, L. M., & Bishop, B. J. 2004, "Random error in willingness to pay measurement: A multiple indicators, latent variable approach to the reliability of contingent values", *Journal of Economic Psychology*, vol. 25, no. 1, pp. 41-59.

Kahneman, D. & Knetsch, J. L. 1992, "Valuing Public-Goods - the Purchase of Moral Satisfaction", *Journal of Environmental Economics and Management*, vol. 22, no. 1, pp. 57-70.

Kahneman, D., Slovic, P., & Tversky, A. 1982, *Judgment under uncertainty: Heuristics and biases* Cambridge University Press, Cambridge, U.K.

Kahneman, D. & Tversky, A. 2000, *Choices, Values and Frames* Cambridge University Press, New York.

Keane, M. P. 1997, "Current Issues i Discrete Choice Modeling", *Marketing Letters*, vol. 8, no. 3, pp. 307-322.

Keeney, R. L. & Raiffa, H. 1976, *Decisions with Multiple Objectives - Preferences and Value Tradeoffs* John Wiley & Sons, New York.

Krinsky, I. & Robb, A. L. 1986, "On Approximating the Statistical Properties of Elasticities", *Review of Economics and Statistics*, vol. 68, no. 4, pp. 715-719.

Krutilla, J. V. 1967, "Conservation Reconsidered", *The American Economic Review,* vol. 57, no. 4, pp. 777-786.

Kuhfeld, W. F. 2003, *Marketing Research Methods in SAS Experimental Design, Choice, Conjoint, and Graphical Techniques ...* SAS Institute Inc., Cary.

Kuhfeld, W. F., Tobias, R. D., & Garatt, M. 1994, *Efficient Experimental Design with Marketing Research Applications* SAS Institute,TS-650C.

Lancaster, K. J. 1966, "A new approach to consumer theory", *Journal of Political Economy*, vol. 74, no. 2, pp. 132-157.

Lancsar, E. & Savage, E. 2004, "Deriving welfare measures from discrete choice experiments: inconsistency between current methods and random utility and welfare theory
1", *Health Economics*, vol. 13, no. 9, pp. 901-907.

List, J. A. & Gallet, C. A. 2001, "What experimental protocol influence disparities between actual and hypothetical stated values?", *Environmental & Resource Economics*, vol. 20, no. 3, pp. 241-254.

Lloyd, A. J. 2003, "Threats to the estimation of benefit: are preference elicitation methods accurate?", *Health Economics*, vol. 12, no. 5, pp. 393-402.

Louviere, J. 2000, *Why stated preference discrete chioice modelling is NOT conjoint analysis*, Memetrics White Paper.

Louviere, J., Hensher, D. A., & Swait, J. 2000, *Stated Choice Methods, analysis and application* Cambride University Press, U.K.

Louviere, J., Street, D., Carson, R., Ainslie, A., Deshazo, J. R., Cameron, T., Hensher, D., Kohn, R., & Marley, T. 2002, "Dissecting the random component of utility", *Marketing Letters*, vol. 13, no. 3, pp. 177-193.

Louviere, J. J. 2001a, "Choice Experiments: an overview of Concepts and Issues" in *The choice Modelling Approach to Environmental Valuation*, J. Bennett & R. Blamey, eds., Edward Elgar Publishing Limited, Cheltenham, UK, pp. 13-36.

Louviere, J. J. 2001b, "What if consumer experiments impact variances as well as means? Response variability as a behavioral phenomenon", *Journal of consumer research*, vol. 28, no. 3, pp. 506-511.

Louviere, J. J. & Woodworth, G. 1983, "Design and Analysis of Simulated Consumer Choice Or Allocation Experiments - An Approach Based on Aggregate Data", *Journal of Marketing Research*, vol. 20, no. 4, pp. 350-367.

Luce, R. D. 1959, *Individual Choice Behavoir: A Theoretical Analysis* Wiley, New York.

Luce, R. D. & Tukey, J. W. 1964, "Simultaneous Conjoint-Measurement - A New Type of Fundamental Measurement", *Journal of Mathematical Psychology*, vol. 1, no. 1, pp. 1-27.

Lusk, J. L. 2003, "Effects of cheap talk on consumer willingness-to-pay for golden rice", *American Journal of Agricultural Economics*, vol. 85, no. 4, pp. 840-856.

Lusk, J. L. & Schroeder, T. C. 2004, "Are choice experiments incentive compatible? A test with quality differentiated beef steaks", *American Journal of Agricultural Economics*, vol. 86, no. 2, pp. 467-482.

Maddala, T., Phillips, K. A., & Johnson, F. R. 2003, "An experiment on simplifying conjoint analysis designs for measuring preferences", *Health Economics*, vol. 12, no. 12, pp. 1035-1047.

Manski, C. H. 1977, "The Structure of Random Utility Models", *Theory and Decision*, vol. 8, pp. 229-254.

Mark, T. L. & Swait, J. 2004, "Using stated preference and revealed preference modeling to evaluate prescribing decisions", *Health Economics*, vol. 13, no. 6, pp. 563-573.

McFadden, D. 1974, "Conditional logit analysis of qualitative choice behaviour," in *Frontiers of Econometrics*, P. Zarembka, ed., Academic Press, London, U.K., pp. 105-142.

McFadden, D. 1980, "Econometric Models for Probabilistic Choice Among Products", *The Journal of Business*, vol. 53, no. 3, pp. 13-29.

McFadden, D. 1986, "The choice theory approach to market research", *Marketing Science*, vol. 5, no. 4, pp. 275-297.

McFadden, D. 1999, "Rationality for economists?", *Journal of Risk and Uncertainty*, vol. 19, no. 1-3, pp. 73-105.

McFadden, D. 2001, "Economic Choices", *AMERICAN ECONOMIC REVIEW*, vol. 91, no. 3, pp. 351-378.

McFadden, D. & Train, K. 2000, "Mixed MNL models for discrete response", *Journal of Applied Econometrics*, vol. 15, no. 5, pp. 447-470.

McIntosh, E. & Ryan, M. 2002, "Using discrete choice experiments to derive welfare estimates for the provision of elective surgery: Implications of discontinuous preferences", *Journal of Economic Psychology*, vol. 23, no. 3, pp. 367-382.

Mitchell, R. M. & Carson, R. T. 1889, *Using Surveys to Value Public Goods: The Contingent Valuation Method* Resources for the Future, Washington, D.C.

O'Brien, B. & Gafni, A. 1996, "When do the "Dollars" make sense? Toward a conceptual framework for contingent valuation studies in health care", *Medical Decision Making*, vol. 16, no. 3, pp. 288-299.

Ohler, T., Le, A., Louviere, J., & Swait, J. 2000, "Attribute Range Effects in Binary Response Tasks", *Marketing Letters*, vol. 11, no. 3, pp. 249-260.

Olsen, J. A. 1997, "Aiding priority setting in health care: Is there a role for the contingent valuation method?", *Health Economics*, vol. 6, no. 6, pp. 603-612.

Olsen, J. A., Donaldson, C., & Pereira, J. 2004, "The insensitivity of 'willingness-to-pay' to the size of the good: New evidence for health care", *Journal of Economic Psychology*, vol. 25, no. 4, pp. 445-460.

Olsen, J. A. & Smith, R. D. 2001, "Theory versus practice: A review of 'willingness-to-pay' in health and health care", *Health Economics*, vol. 10, no. 1, pp. 39-52.

Phillips, K. A., Johnson, F. R., & Maddala, T. 2002, "Measuring what people value: A comparison of "attitude" and "preference" surveys", *Health Services Research*, vol. 37, no. 6, pp. 1659-1679.

Ratcliffe, J. & Buxton, M. 1999, "Patients' preferences regarding the process and outcomes of life-saving technology - An application of conjoint analysis to liver transplantation", *International Journal of Technology Assessment in Health Care*, vol. 15, no. 2, pp. 340-351.

Ratcliffe, J. & Longworth, L. 2002, "Investigating the structural reliability of a discrete choice experiment within health technology assessment", *International Journal of Technology Assessment in Health Care*, vol. 18, no. 1, pp. 139-144.

Revelt, D. & Train, K. 1998, "Mixed logit with repeated choices: Households' choices of appliance efficiency level", *Review of Economics and Statistics*, vol. 80, no. 4, pp. 647-657.

Rolfe, J., Bennett, J., & Louviere, J. 2002, "Stated values and reminders of substitute goods: Testing for framing effects with choice modelling", *Australian Journal of Agricultural and Resource Economics*, vol. 46, no. 1, pp. 1-20.

Rosen, S. 1974, "Hedonic Prices and Implicit Markets - Product Differentiation in Pure Competition", *Journal of Political Economy*, vol. 82, no. 1, pp. 34-55.

Ryan, M. 1994, "Agency in Health-Care - Lessons for Economists from Sociologists", *American Journal of Economics and Sociology*, vol. 53, no. 2, pp. 207-217.

Ryan, M. 1996, "Using willingness to pay to assess the benefits of assisted reproductive techniques", *Health Economics*, vol. 5, no. 6, pp. 543-558.

Ryan, M. 1999a, "A role for conjoint analysis in technology assesment in health care?", *International Journal of Technology Assessment in Health Care*, vol. 15, no. 3, pp. 443-457.

Ryan, M. 1999b, "Using conjoint analysis to take account of patient preferences and go beyond health outcomes: an application to in vitro fertilisation", *Social Science & Medicine*, vol. 48, no. 4, pp. 535-546.

Ryan, M. 2004a, "A comparison of stated preference methods for estimating monetary values", *Health Economics*, vol. 13, no. 3, pp. 291-296.

Ryan, M. 2004b, "Deriving welfare measures in discrete choice experiments: a comment to Lancsar and Savage (1)", *Health Economics*, vol. 13, no. 9, pp. 909-912.

Ryan, M. 2004c, "Discrete choice experiments in health care - NICE should consider using them for patient centred evaluations of technologies", *British Medical Journal*, vol. 328, no. 7436, pp. 360-361.

Ryan, M. & Gerard, K. 2003, "Using discrete choice experiments to value health care programmes: current practice and future research reflections", *Appl.Health Econ.Health Policy*, vol. 2, no. 1, pp. 55-64.

Ryan, M. & Hughes, J. 1997, "Using conjoint analysis to assess women's preferences for miscarriage management", *Health Economics*, vol. 6, no. 3, pp. 261-273.

Ryan, M. & San Miguel, F. 2003, "Revisiting the axiom of completeness in health care", *Health Economics*, vol. 12, no. 4, pp. 295-307.

Ryan, M., Scott, D. A., & Donaldson, C. 2004, "Valuing health care using willingness to pay: a comparison of the payment card and dichotomous choice methods", *J.Health Econ.*, vol. 23, no. 2, pp. 237-258.

Ryan, M. & Skatun, D. 2004, "Modelling non-demanders in choice experiments", *Health Economics*, vol. 13, no. 4, pp. 397-402.

Ryan, M. & Ubach, C. 2003, "Testing for an experience endowment effect in health care", *Applied Economics Letters*, vol. 10, no. 7, pp. 407-410.

Ryan, M. & Wordsworth, S. 2000, "Sensitivity of willingness to pay estimates to the level of attributes in discrete choice experiments", *Scottish Journal of Political Economy*, vol. 47, no. 5, pp. 504-524.

Saelensminde, K. 2001, "Inconsistent choices in Stated Choice data - Use of the logit scaling approach to handle resulting variance increases", *Transportation*, vol. 28, no. 3, pp. 269-296.

Salkeld, G., Ryan, M., & Short, L. 2000, "The veil of experience. Do consumers prefer what they know best?", *Health Economics*, vol. 9, no. 3, pp. 267-270.

Salkeld, G., Solomon, M., Short, L., Ryan, M., & Ward, J. E. 2003, "Evidence-based consumer choice: a case study in colorectal cancer screening", *Aust.N.Z.J.Public Health*, vol. 27, no. 4, pp. 449-455.

Samnaliev, M., Stevens, M., & More, T. A Comparison of Cheap Talk and Alternatie Certainty Calibration Techniques in Contingent Valuation. Working Paper No. 2003-11. University of Massachusetts Amherst. 2003.

San Miguel, F., Ryan, M., & McIntosh, E. 2000, "Applying conjoint analysis in economic evaluations: an application to menorrhagia", *Applied Economics*, vol. 32, no. 7, pp. 823-833.

San Miguel, F., Ryan, M., & Scott, A. 2002, "Are preferences stable? The case of health care", *Journal of Economic Behavior & Organization*, vol. 48, no. 1, pp. 1-14.

Schumacher, E. J. & Whitehead, J. C. 2000, "The production of health and the valuation of medical inputs in wage-amenity models", *Soc.Sci.Med.*, vol. 50, no. 4, pp. 507-515.

Scott, A. 2002, "Identifying and analysing dominant preferences in discrete choice experiments: An application in health care", *Journal of Economic Psychology*, vol. 23, no. 3, pp. 383-398.

Scott, A. & Vick, S. 1999, "Patients, doctors and contracts: An application of principal-agent theory to the doctor-patient relationship", *Scottish Journal of Political Economy*, vol. 46, no. 2, pp. 111-134.

Scott, A., Watson, M. S., & Ross, S. 2003, "Eliciting preferences of the community for out of hours care provided by general practitioners: a stated preference discrete choice experiment", *Soc.Sci.Med.*, vol. 56, no. 4, pp. 803-814.

Severin, V., Louviere, J. J., & Finn, A. 2001, "The stability of retail shopping choices over time and across countries", *Journal of Retailing*, vol. 77, no. 2, pp. 185-202.

Shiell, A. & Gold, L. 2003, "If the price is right: vagueness and values clarification in contingent valuation", *Health Economics*, vol. 12, no. 11, pp. 909-919.

Silva, J. M. C. S. 2004, "Deriving welfare measures in discrete choice experiments: a comment to Lancsar and Savage (2)", *Health Economics*, vol. 13, no. 9, pp. 913-918.

Simon, H. A. 1982, *Models of Bounded Rationality* The MIT Press, Cambridge.

Skjoldborg, U. S. & Gyrd-Hansen, D. 2003, "Conjoint analysis. The cost variable: an Achilles' heel?", *Health Economics*, vol. 12, no. 6, pp. 479-491

Small, K. A. & Rosen, H. S. 1981, "Applied Welfare Economics with Discrete Choice Models", *Econometrica*, vol. 49, no. 1, pp. 105-130.

Smith, R. D. 2003, "Construction of the contingent valuation market in health care: a critical assessment", *Health Economics*, vol. 12, no. 8, pp. 609-628.

Swait, J. & Adamowicz, W. 2001a, "Choice environment, market complexity, and consumer behavior: A theoretical and empirical approach for incorporating decision complexity into models of consumer choice", *Organizational Behavior and Human Decision Processes*, vol. 86, no. 2, pp. 141-167.

Swait, J. & Adamowicz, W. 2001b, "The influence of Task Complexity on Consumer Choice: A Latent Class Model of Decision Strategy Switching", *Journal of consumer research*, vol. 28, no. June, pp. 135-148.

Swait, J., Adamowicz, W., Hanemann, M., Diederich, A., Krosnick, J., Layton, D., Provencher, W., Schkade, D., & Tourangeau, R. 2002, "Context dependence and aggregation in disaggregate choice analysis", *Marketing Letters*, vol. 13, no. 3, pp. 195-205.

Swait, J., Adamowicz, W., & van Bueren, M. 2004, "Choice and temporal welfare impacts: incorporating history into discrete choice models", *Journal of Environmental Economics and Management*, vol. 47, no. 1, pp. 94-116.

Swait, J. & Louviere, J. 1993, "The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit-Model", *Journal of Marketing Research*, vol. 30, no. 3, pp. 305-314.

Taylor, S. & Armour, C. 2003, "Consumer preference for dinoprostone vaginal gel using stated preference discrete choice modelling", *Pharmacoeconomics*, vol. 21, no. 10, pp. 721-735.

Thurstone, L. L. 1927, "A Law of Comparative Judgment", *Psychological Review*, vol. 34, pp. 273-286.

Train, K. 1993, *Qualitative Choice Analysis: Theory Econometrics, and an Application to Automobile Demand*, 3 edn, The MIT Press, Cambridge, Massachusetts.

Train, K. 2003, *Discrete Choice Methods with Simulation* Cambridge University Press, U.K..

Train, K. & Sonnier, G. Mixed Logit with Bounded Distributions of Partworths. 2003.
Ref Type: Unpublished Work

Train, K. E. 1998, "Recreation demand models with taste differences over people", *Land Economics*, vol. 74, no. 2, pp. 230-239.

Tversky, A. 1972, "Elimination by Aspects - Theory of Choice", *Psychological Review*, vol. 79, no. 4, p. 281-&.

Tversky, A. & Kahneman, D. 1991, "Loss Aversion in Riskless Choice - A Reference-Dependent Model", *Quarterly Journal of Economics*, vol. 106, no. 4, pp. 1039-1061.

Tyrvainen, L. & Miettinen, A. 2000, "Property prices and urban forest amenities", *Journal of Environmental Economics and Management*, vol. 39, no. 2, pp. 205-223.

Ubach, C., Scott, A., French, F., Awramenko, M., & Needham, G. 2003, "What do hospital consultants value about their jobs? A discrete choice experiment", *BMJ*, vol. 326, no. 7404, p. 1432.

van der Pol, M. & Cairns, J. 2001, "Estimating time preferences for health using discrete choice experiments", *Social Science & Medicine*, vol. 52, no. 9, pp. 1459-1470.

Verbeek, M. 2000, *A giude to Modern Econometrics* John Wiley & SOns, Ltd, Chichester.

von Haefen, R. H. 2003, "Incorporating observed choice into the construction of welfare measures from random utility models", *Journal of Environmental Economics and Management*, vol. 45, no. 2, pp. 145-165.

Wainwright, D. M. 2003, "More 'con' than 'joint': problems with the application of conjoint analysis to participatory healthcare decision making", *Critical Public Health*, vol. 13, no. 4, pp. 373-380.

Weisbrod, B. A. 1964, "Collective-Consumption Services of Individual-Consumption Goods", *Quarterly Journal of Economics*, vol. 78, no. 3, pp. 471-477.

Willig, R. D. 1976, "Consumers Surplus Without Apology
9", *American Economic Review*, vol. 66, no. 4, pp. 589-597.

Willis, K. G. & Garrod, G. D. 1991, "An Individual Travel-Cost Method of Evaluating Forest Recreation", *Journal of Agricultural Economics*, vol. 42, no. 1, pp. 33-42.

Wittink, D. R. & Cattin, P. 1989, "Commercial Use of Conjoint-Analysis - An Update
2", *Journal of Marketing*, vol. 53, no. 3, pp. 91-96.

Wooldridge, J. M. 2002, *Econometric analysis of Cross Section and Panel Data* The MIT Press, London.

Zhao, J. H. & Kling, C. L. 2001, "A new explanation for the WTP/WTA disparity", *Economics Letters*, vol. 73, no. 3, pp. 293-300.

Zwerina, K., Huber, J., & Kuhfeld, W. F. 1996, *A General Method for Constructing Efficient Choice Designs*.

Appendix I. Illustrations of welfare measures

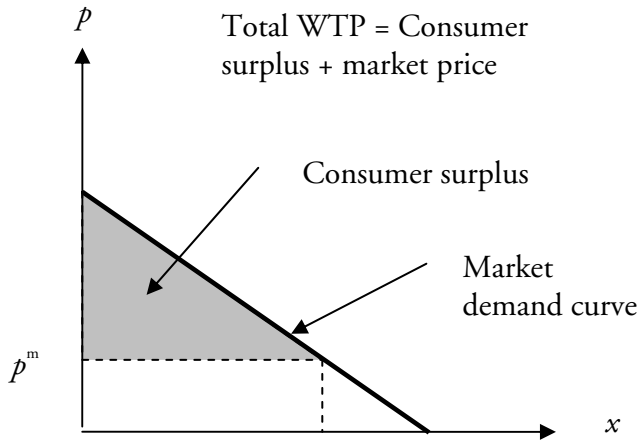

Figure A1: Marshallian demand curve, consumer surplus and WTP

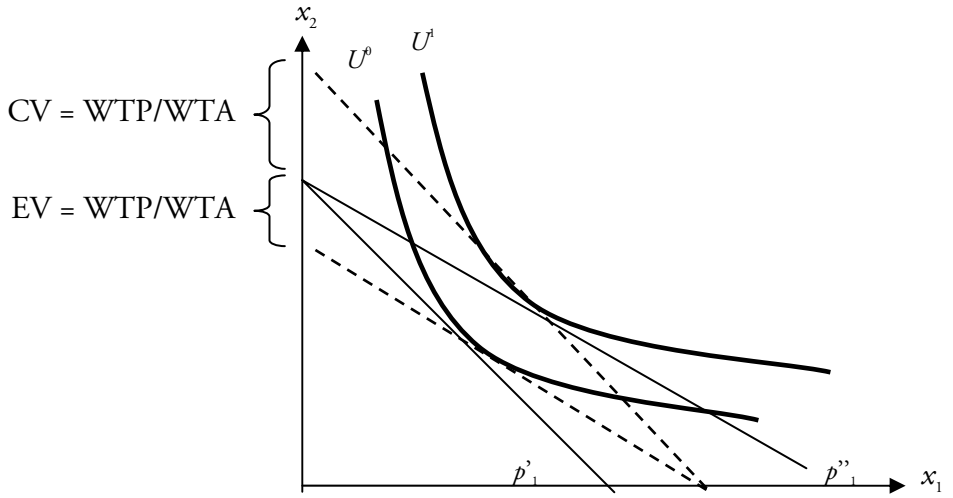

Figure A2: Hicksian welfare measures in a market with two goods (welfare gain due to price decrease: from P' to P'')

CV: $U^0(m, p', x_j) = U^1(m - \mathrm{CV}, p'', x_j)$

EV: $U^0(m + \mathrm{EV}, p', x_j) = U^1(m, p'', x_j)$

Appendix II. Illustrations of coding and data input

Table A1. An example of a DCE question and coding

	Mode 1	Coding	**Mode 2**	Coding	Δ Modes = (mode 2 – mode 1)
Means of transport	Car	D_{car}* =1 (D_{bus} =0)	Bus	D_{bus}* =1 (D_{car} =0)	D_{car} =-1 D_{bus} =1
Duration of transport	20 minutes	=20	25 minutes	=25	=5
Walk to/from transportation	2 minutes	=2	10 minutes	=10	=8
Price	kr. 50	=50	kr. 20	=20	=-30
I choose the transportation mode	☐	Input as differences: =0 (mode 1 is chosen) Input both alternatives: = 0 not chosen, =1 chosen	☐	Input as differences: =0 (mode 1 is chosen) Input both alternatives: = 0 not chosen, =1 chosen	

* Dummy variables, coded (=1) when present and (=0) when not present. Effect coding would have been an alternative.

Table A2. Data input by differences (the first question is that shown in A1 above). This solution can only be applied to DCEs with two alternatives. For data analysis in Stata, binary choice tasks have to be coded as differences.

Id	Question	Answer	Δcar	Δbus	Δduration	Δwalk	Δprice
1	1	0	-1	1	5	8	-30
1	2	1	1	-1	10	-5	40
1	3	1	0	0	7	-7	10
2	1	0	-1	1	5	8	-30
2	2	0	1	-1	10	-5	40
2	3	1	0	0	7	-7	10
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.

Table A3. Data input by each alternative (the first question is that shown in A1 above). This is always the procedure when the number of alternatives exceeds two.

id	Question	Alternative (mode)	Answer	Car	Bus	Duration	Walk	Price
1	1	1	1	1	0	20	2	50
1	1	2	0	0	1	25	10	20
1	2	1	0	0	1	10	10	40
1	2	2	1	1	0	20	5	80
1	3	1	0	0	1	7	14	5
1	3	2	1	0	0	14	7	15
2	1	1	1	1	0	20	2	50
2	1	2	0	0	1	25	10	20
2	2	1	1	0	1	10	10	40
2	2	2	0	1	0	20	5	80
2	3	1	0	0	1	7	14	5
2	3	2	1	0	0	14	7	15
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.

As can be seen from both examples, respondent 1 (id=1) chooses alternative 1 whereas respondent 2 (id=2) chooses alternative 2.

Note: For programming in Gauss, each alternative needs to be specified separately, however in one line (row).

Appendix III. An example of a regression analysis

Table A4. A random effects probit: a study of students' preferences for their future jobs

	Coefficient (p)	**Marginal WTP**
Wage (D.kr.)	0,00025 (0,00)	
Working hours	-0,15300 (0,00)	599 (= -0,15300/-0,00025)
Private versus public sector (0=private, 1=public)	0,15462 (0,01)	-605 (=0,15462/0,00025)
Increased flexibility (working at home with computer, e-mail and phone at disposal)	0,51099 (0,00)	-2000 (=0,51099/0,00025)
N=749 Wald chi (p)=273,7 (0,00) LogL=-326,066 μ=0,0009		

The results are obtained from a DCE of students' preferences for their future jobs. The middle column shows the coefficients from the random effects model, while the right hand column shows the estimated marginal rates of substitution (MRS) between wage and the other job characteristics, i.e. WTP. MRS has been estimated by dividing the coefficient for two attributes. The results show that all coefficients are highly significant. The wage attribute is positive, which implies a positive utility of 0.00025 for earning one extra Danish Kr. The sign for working hours is negative, which implies a negative utility of -0.153 for working one extra hour. The students also report positive utility for working in the public sector and for greater work flexibility. The WTP estimates in the right hand column indicate that the students are willing to work one extra hour if their wage is increased by 599 D.kr. Moreover, they are willing to accept a reduction in salary of 605 D.kr. for a job in the public sector and of 2000 D.kr. to have greater job flexibility. Random effects models make it possible to take into consideration the person-specific (within) variation that might exist in the data. In the above example, this error term is estimated to be 0.0009, which means that 0.9% of unobserved utility is due to within-variation.