UTS:CenSoC
Centre for the Study of Choice

UNIVERSITY OF TECHNOLOGY SYDNEY

Centre for the Study of Choice
University of Technology, Sydney
Level 4, 645 Harris Street
Ultimo NSW 2007
PO Box 123
T: +61 2 9514 9799
F: +61 2 9514 9897
www.censoc.uts.edu.au
UTS CRICOS PROVIDER CODE 00099F

Centre for the Study of Choice (CenSoC)

Working Paper Series

No. 10-001

'Using Conjoint Analysis to Estimate Health State Values for Cost-Utility Analysis: Issues to Consider'

By Terry N. Flynn

January, 2010

Using conjoint analysis to estimate health state values for Cost-Utility Analysis: Issues to Consider

Terry N Flynn PhD
CenSoC - Centre for the Study of Choice, University of Technology Sydney
645 Harris St Ultimo NSW 2007
Mailing address: PO Box 123 Broadway NSW 2007 Australia

Tel (direct) +61 2 9514 9804
Fax +61 2 9514 9897
E-mail: terry.flynn@uts.edu.au

Abstract

Researchers are increasingly using a variety of conjoint tasks, including ranking, discrete choice experiments and best-worst scaling studies to estimate health state (QALY) values for use in cost-utility analysis. There are a number of serious conceptual, theoretical and empirical difficulties in conducting such studies that have received insufficient attention in the health literature. These include proper modelling of the variance scale factor, which potentially influential papers have failed to do. Recent research has also cast serious doubt on the properties of random effects models (such as the mixed logit) in choice experiments. This paper details these concerns and offers recommendations on the conduct of 21st century QALY valuation exercises that propose to use any tasks that rely on discrete choices.

Using conjoint analysis to estimate health state values for Cost-Utility Analysis: Issues to Consider

When estimating health state values for cost-utility analysis (CUA), QALY valuation exercises have typically used choice tasks that rely on cardinality of preferences: respondents are assumed to be able to provide numerical responses that indicate strength of preference. Serious issues have been identified with the theoretical and empirical properties of these Standard Gamble and Time Trade-Off tasks.[1, 2] Furthermore, the tasks are perceived to be difficult, typically requiring administration by interview. As a result there is growing interest in tasks that only require the weaker assumption of ordinality of preferences: respondents are only required to make discrete choices (A or B). These choice tasks include ranking studies, discrete choice experiments (called choice-based conjoint analyses in North America) and best-worst scaling (BWS) studies.

Generally, it is advantageous to use a choice task that requires fewer, or weaker, assumptions about human decision-making. However, there is no such thing as a free lunch and this advantage comes at a cost: correct interpretation of discrete choice data can be extremely difficult. There is a danger that important national valuation exercises to provide decision makers with 21st century QALY values will be seriously flawed, due to failure to address key issues that are on the research frontier. The purpose of this paper is to set out the key theoretical, conceptual and empirical issues that all practitioners of CUA should be aware of if they intend to conduct conjoint studies, such as discrete choice experiments, ranking studies or best-worst scaling studies. Some of these issues are the subject of current work; crucially, the first solutions have been put forward in the marketing and environmental economics (rather than health) literature. This paper aims to introduce these issues to a health audience and to discuss their implications. The aim is *not* to argue for the wholesale

rejection of the standard gamble and time trade-off methods used to implement past valuation exercises. Indeed, study designs that nest those tasks in a wider choice experiment will be discussed. Rather, the paper will outline the issues in using ordinal tasks before discussing the implications of their use in QALY valuation exercises.

1. Ordinal tasks

The use of choice tasks that rely only on ordinality, not cardinality, is well-established in fields such as mathematical psychology.[3] In terms of valuing instruments (estimating the 'scoring') for use in CUA, three principal ordinal tasks have been administered. Ranking exercises were conducted as a warm up to studies such as the UK MVH study for EQ-5D,[4] discrete choice experiments have been used for the OSCA social care study,[5] and are under consideration for EQ-5D-5L, whilst best-worst scaling has been used for ICECAP-O.[6]

The assumption of ordinality requires that respondents be able to indicate only a preference (A is preferred to B) but not the strength of preference (how much A is preferred to B, or how much of B is required to ensure indifference between it and A). This use of a discrete rather than continuous outcome necessitates the use of limited dependent variable models, such as logistic or probit regression. Unfortunately, many of the skills acquired when using these methods in fields such as epidemiology and health economics are *not* transferable. This is because epidemiologists and many applied health researchers are interested in making inference about probabilities (such as odds ratios). However, the researcher interested in obtaining values for use in CUA aims to use the probabilities (the observed choice frequencies), to make inferences about the underlying latent scale (health or quality of life in a non-welfarist framework, utility in a welfarist one). Unfortunately, there is a critical limitation of all limited dependent variable models: they perfectly confound

estimates of the mean and variance (typically discussed in terms of its inverse – the variance scale factor or simply 'scale factor') on the underlying latent scale.[7] A large point estimate from any of these models might represent a large mean (preference), a small variance (high consistency of responses) or any of an infinite number of combinations of the two. Since all standard statistical packages arbitrarily set the variance scale factor to be one (to enable identification), there is a temptation for the researcher not to model variance heterogeneity explicitly. Unfortunately, as will be discussed below, this has extremely severe adverse consequences for interpreting discrete choice data for use in CUA. Dealing with scale is easiest when it is conceptualised and modelled in the theoretical framework underpinning all of the three types of discrete choice models considered in this paper: random utility theory (RUT).

2. Random utility theory

Random utility theory (RUT) was set out and operationalised by Thurstone and McFadden.[8, 9] Indeed McFadden won the Nobel Prize for Economics in 2000 in part for proving that multinomial (conditional) logistic (MNL) regression can be used to estimate the parameters of a random utility choice model, subject to certain assumptions. (He acknowledged Tony Marley in his Nobel lecture for proving this independently). RUT is both simple and difficult. It is simple in that conceptually it states that the difference in utility between two objects, A and B, is proportional to the frequency that one is chosen over the other. It is difficult due to the particular *ceteris paribus* statement required to operationalise it: namely that the variance of the random utility term is constant. This statement is frequently ignored, or if not, is not understood in terms of the extremely serious implications it can have for the correct interpretation of the parameters of the regression model. Differences in the variance of the random utility term can arise from any of a number of reasons, but perhaps the easiest to relate to is a respondent's choice consistency. A respondent who is very certain of his/her

preferences, choosing very consistently across all choice occasions/sets, exhibits a low variance of the random utility term (which has mean zero). On the other hand a respondent who is rather uncertain of his/her preferences, choosing relatively inconsistently across choice occasions, exhibits a high variance.

The likelihood of different variances is fairly well understood in other branches of economics and in the econometrics literature.[10] However, this acknowledgement is largely restricted to the need to adjust for differences when making *between* study comparisons: this may be due to the fact that the economist's 'trick' of dividing estimates of all other attributes by that of the price attribute to estimate willingness to pay (WTP) also has the statistical advantage of cancelling the confounded variance scale factor from numerator and denominator. However, the problem of mean-variance confounding is equally pertinent *within* a given study.[11] Indeed it is even more serious since any differences by attribute in the random utility term cause the 'WTP trick' to fail: if there isn't a common variance scale factor then dividing through by the price coefficient doesn't cancel it. The welfare estimates are then biased. The overall problem can be summed up in the following statement:

If the variance of the random utility term is not constant across all attributes, across all choices and across all respondents then the mean preference estimates are inflated/deflated differentially; in short, the regression estimates from a traditional conditional logit model are biased, in unknown direction and with unknown magnitude.

The above represents a theoretical problem. Pertinent questions are "how likely is it to happen in real life?" and "what methods can be used to deal with the problem?"

3. The importance of variance heterogeneity

Some DCE practitioners will claim that the above is unduly pessimistic, perhaps even scare-mongering. In discussing the popular mixed logit (MIXL or MMNL) model, which introduces random effects, McFadden and Train state "...any discrete choice model derived from random utility maximization has choice probabilities that can be approximated as closely as one pleases by a MMNL".[12] Yet in practice a multivariate normal distribution is usually used to model the attribute coefficients.[13] This is an assumption and is not usually tested in health applications. This may, again, be due to an over-reliance upon methods and assumptions made in health that do not carry over to choice models, namely that normal distributions are reasonable. Unfortunately, seminal work in press by Michael Keane, the 2009 winner of the Kenneth Arrow prize for health economics, and colleagues comprehensively shows that this assumption leads to incorrect inference in 100% of the choice model applications they consider (including monte carlo simulations):[13] in short, the mixed logit model is *always* dominated by models that *explicitly* model variance heterogeneity. Furthermore, the discipline in which variance heterogeneity is typically most prevalent is (somewhat unsurprisingly) health: variation in choice consistency is much lower when people are deciding which TV to buy than when they are choosing health insurance plan or treatment.

It should be noted that Fiebig, Keane *et al* do not claim their proposed 'generalised multinomial logistic regression (G-MNL) model is 'the' solution to the above problem. The mean-variance confound means that in many cases, it will be impossible to state that any 'gold standard' method has been used. Instead, researchers should set out clearly (and justify) the assumptions made in their heterogeneity analyses. Given their results, advocates of MIXL, in particular, should have to demonstrate this to referees and editors. Section 8 will discuss further why designers of QALY valuation exercises cannot ignore this issue on grounds of 'only' requiring population average

preferences. However, the next sections will first discuss issues that are specific to each of the different types of conjoint task.

4. Ranking studies

Ranking as a choice task has a long history, with seminal work conducted by Luce and Marley,[14, 15] and the first empirical study to use the rank ordered logit (ROL) model reported by Beggs and Cardell.[16] More recently, it has been proposed that a simple ranking of all relevant health states, together with the death state, might provide the QALY values required for CUA.[17, 18] However, contrary to statements made in those papers, the logistic regression estimates from ranking studies are *not utilities on that scale*. There was no acknowledgement of the confounded scale factor in the definition of the log odds ratio given in McCabe *et al* and the variance scale factor was not mentioned by either paper. It is worrying that these two potentially influential papers in the health economics literature appear not to understand the mean-variance confound.

A second problem with the ranking models proposed in both papers can be set out as follows: if any individual respondent makes choices deterministically (rather than probabilistically) – in other words, chooses without error – then they violate the assumptions of RUT and should not be included in any statistical model (such as the MNL and all its generalisations) that relies on RUT.[19] Choices between life and death (when made without varying risk or length of life) *are* made deterministically by some people, particularly those who believe it is not for humans to decide that death is preferable to a living state, no matter how bad it is.

Anecdotally this author has heard researchers state they can ‘fix’ this problem, typically by presenting a health state that is so terrible that the respondent finally chooses death in preference to

it. Yet, if such a state (say, characterised by the worst possible torture continuing for the rest of your life) is not part of the descriptive system then one is using a totally unrealistic scenario to escape a problem that remains: there are people for whom there is no realistic state of the world in which they would prefer death; there is no RU term, they choose deterministically, life over death.

The ranking studies of McCabe *et al* and Salomon cannot be 'fixed' to produce the correct QALY trade-offs. This is because length of life, a key factor (indeed the numeraire) is not present. Consider the figure below, which explains the concept underlying the TTO. The two rectangles of equal area: they represent the two quantity-quality combinations between which a respondent is indifferent. When the period of impaired health is varied (from the 50 years in the example), if the utility of life duration is constant, then the various health states will maintain their relative positions on a line that radiates from the origin to the top right corner of the graph.

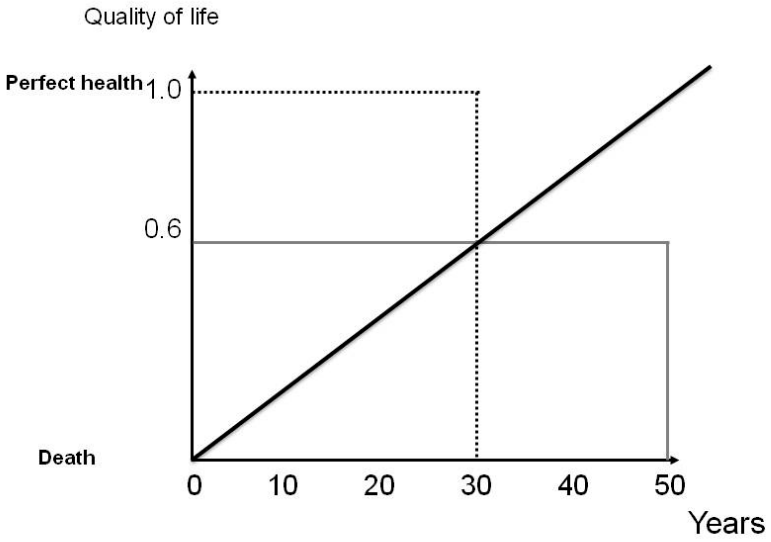

Figure 1. Representation of a completed Time Trade-off exercise question

If the utility of life duration varies then the positions of the health states relative to the origin changes; for example when the TTO is asked with reference to a one year timeframe it is likely that, *relative to the origin*, all the intersection point lies much closer to the top right point – which is now the point (10,1) rather than (50,1) – as individuals are reluctant to give up any life. The ranking task appears to estimate the positions of the various health states when the line is rotated so it coincides with the positive y-axis. However, the choice is between various impaired lives of a given (but often unstated) length; death can be properly described qualitatively but it is absolute in terms of length. The choice between it and any living state is *not* the same as that between living states.

There are two key analysis issues the researcher should acknowledge when analysing ranking data. First, (s)he should test whether the variance scale factor differs at different depths of ranking (for instance, whether respondents are less sure of their middle rankings than their top and bottom ones).[20] Second, after adjusting for any such differences in variance scale factor, whether the same estimates are obtained from the various ranking depths: for instance, do the bottom, middle and top rankings all estimate the same, common, utility function? Key papers in the marketing and mainstream economics literature have found that the scale factor typically varies by ranking depth and, more worryingly, even after adjusting for this the mean estimates at different depths also vary.[21] This suggests that poor information is being obtained (particularly from the middle rankings) and is one of the reasons why best-worst scaling (described in Section 6) was invented by Jordan Louviere.[22]

In conducting a ranking study care must be taken over the order in which the questions are asked. The commonly implemented ROL model available in many statistical packages assumes a very particular psychological model of choice, namely most preferred from *n* states, most preferred from *n-1* when the chosen state is removed, etc through to choice from the final 2 states. The asymmetry

in the EV1 (gumbel) distribution assumed for the random utility term means that the result will typically differ slightly from that obtained if one implemented a ROL model which went from least preferred up to most preferred.[20] So, question format matters, although if the ordering of the health state attribute levels changes depending on which model is estimated, this raises questions as to the stability of the model in the first place. Nevertheless, given that many statistical ranking models assume a particular psychological model (in other words, how the ranking was obtained), researchers should pay close attention to how they ask ranking questions. They can be sure to use the appropriate statistical model if they have used either interviews or the web to structure a ranking task (which has the added benefit of making it easier). This is a particularly strong reason why any 'free' ranking methods should be discouraged: one has no idea which psychological model of ranking respondents have used (and therefore which statistical model is appropriate for that); one has done nothing to make the task simple for respondents – indeed one has made it as difficult as it could possibly be for them. The web now offers researchers the opportunity to easily and cost-effectively control the psychological model used by deleting the chosen option each time. Researchers at CenSoC, UTS, now routinely obtain full rankings of items from respondents by asking questions in a particular order: best, worst, next best, next worst, etc.[23] 'Best' and 'worst' are likely to be easiest: it is well established that humans find identification of extremes easy[24] and these two choices are likely to exhibit large (and probably similar) variance scale factors (i.e. small variances). Doing repeated rounds of best-worst in this fashion is likely to be the easiest way to obtain the full ranking whilst keeping the variance scale factor constant within each round, although this is an empirical issue.

Given the results of Fiebig *et al*, introducing random effects to model heterogeneity in ranking data is unlikely to be a fruitful solution. MIXL can certainly not be considered to be a gold standard. If any gold standard exists it is the empirical distribution of preferences obtained by aggregating the utility

functions of all respondents (each one estimated at the level of the individual respondent). This is the subject of current work.[23]

5. Discrete choice experiments

DCEs are explicitly rooted in RUT and, in general, published papers in health have been better at acknowledging these theoretical underpinnings.[25, 26] Many of the key issues in designing and analysing a DCE have already been set out for health economists.[26] Therefore this section will concentrate on those that are specific to CUA studies.

5.1. The 'ideal' DCE

It is important to note that there are a number of ways of estimating QALY tariffs from a DCE. The design of the DCE will be influenced by a number of factors including the resources available and the theoretical model assumed to underpin QALY values. The first DCE considered here might be considered the 'theoretically correct' one. Here 'correct' means the DCE producing estimates that one *would have obtained* from one's preferred cardinal task (primarily SG or TTO) if all respondents experienced no difficulty in iterating to their point of indifference. The DCE would essentially ask respondents to choose between alternative quality-quantity profiles, as in TTO or SG, but would not be asked to do iterative exercises to reach a point of indifference. The choices on offer would be pairs of lives if one wished to maintain a close link with the TTO or SG. Thus a TTO type task would present a shorter life in full health and a longer life in impaired health and the respondent merely indicates which is most preferred. A SG type task would offer a decision with a given risk of death versus the option of accepting an impaired life with certainty.

The options presented to respondents in this study are easy to set up. The statistical design and analysis are potentially problematic and have (large) budgetary implications in terms of sample size. The QALY model is multiplicative: quality x quantity is the maximand. Therefore, a multiplicative DCE model is required, not the additive main effects model of the type used in the vast majority of health applications. In particular, the length of life attribute must be interacted with every quality of life attribute level (5 x 3 = 15 in the case of EQ-5D) and the main effects of the attribute levels must be constrained to be zero (since quality of life without the existence of life itself is nonsensical). The statistical design required here is one that is capable of estimating (at least) those two-way interactions. Estimating *all* two-way interactions requires a *Resolution 5* design which can be a non-trivial fraction of the full factorial (sometimes one half). However, not all two-way interactions may be necessary for a QALY valuation exercise. If (and only if) the researcher is confident that the only two-way interactions that are non-zero are the quantity-quality ones, then a smaller design is possible. It should be noted that if the researcher uses such a smaller design and the omitted interactions are not, in fact, zero then the estimates will be biased:[25] smaller designs are obtained by 'aliasing' (deliberately confounding) higher-order interactions with the lower order ones and main effects that the researcher is estimating.

Large designs immediately run into problems that respondents cannot reasonably be expected to answer all choices. However, in other disciplines this is usually solved by randomising respondents to different subsets, by *blocking the sample*.[25] Ideally one would ask all respondents to answer a common main effects design (to add power to any tests for heterogeneity) and a subset of the remaining interactions design. This clearly has implications for the cost of the valuation exercise, if statistical power is to be preserved.

Advantages of this design include the fact that the utility of life duration can be estimated for various lengths of life. It has been found that respondents are generally less willing to sacrifice length of life to improve quality of life as the amount of time on offer reduces: life itself becomes 'more precious'.[27] Existing QALY tariffs assume a constant utility of life duration since the valuation exercise used a fixed 10 year time horizon. The DCE here can potentially allow this to vary, giving policymakers more information they can use in decisions involving shorter or longer time horizons.

5.2. The smaller (cheaper) DCE

The design above essentially replicates a SG or TTO study without requiring the respondent to iterate to a point of indifference. However, such a study would be expensive to run and it is not clear that policy makers are willing to fund such a study until ordinal task based studies prove their worth. Therefore an intermediate study would be one that is similar to that already conducted for EQ-5D and other instruments, namely a ranking exercise. Such a study would ask respondents to choose their most preferred health state from sets of two or more. It is important to note that, as for all such studies, the researcher must properly protect against outside factors affecting respondents differentially. For example, the length of life must be standardised across all respondents. This type of study, by not varying length of life, potentially requires a much smaller statistical design: if all interactions are zero then one can use a main effects design (18 choice sets for the EQ-5D). Using modulo arithmetic techniques designed by Street & Burgess one can ensure an optimally efficient design – one which does not effectively 'throw away' respondents.[28] At its simplest, such a study would present respondents with 18 pairs of EQ-5D states, asking them to choose their most preferred in each instance. However, Street and Burgess note the potential problems specific to the EQ-5D caused by unrealistic scenarios: a state involving being confined to bed but having no difficulties

with usual activities makes no sense. It is likely that the descriptive system for the EQ-5D-5L will change the bottom level of mobility so as to avoid such problems.

The main problem with this smaller design is that the resulting estimates are anchored to one living health state, not to death.[19] This is because the length of life is fixed, not experimentally varied as it is in the larger design. One *cannot* escape this limitation without asking some question(s) that estimates the quality/quantity trade-off: it has never been mathematically proven that the distance on the latent scale between a given (e.g. pits) health state and death is equal to this parameter (and it is trivial to construct hypothetical examples where it is not). Thus, even if all respondents choose probabilistically between death and living states (i.e. conform to RUT) it is yet to be proved that one would obtain QALY values from this (or a Salomon/McCabe *et al*) ranking task. Any argument that the estimates from such a task seem to agree with those from a TTO or SG is misplaced for two reasons: first, as noted above, theoretically these are two different parameters one is estimating, so the fact they agree may be pure chance; second, there is circular logic in play since the ranking studies have been argued to produce estimates that are better than the TTO ones, by way of the weaker assumptions they use!

Researchers will, of course, ask if there is a way to obtain QALY values from the estimates from the type of DCE described above. The OSCA study in social care-related quality of life will be asking some TTO questions to all (or a subset of the) respondents from the main DCE in order to re-anchor them to death. This is clearly not optimal: the quality/quantity trade-off is being estimated only from those respondents who are able to answer the TTO. However, it is likely that these studies will constitute stepping stones to larger DCEs, if and when the values from these ‘second best’ DCEs can be shown to be robust.

6. Best-worst scaling studies

Best-worst scaling (BWS) was invented by Jordan Louviere in the 1980s as a method of eliciting more information from discrete choice questions.[22, 29, 30] It is rooted in the same random utility framework that underpins DCEs and ranking studies and represents a 'half-way house' between the two: One obtains more information than a DCE and ideally enough to estimate good individual-level utility functions without burdening respondents with providing a full ranking 'in one go'.

BWS has been used to estimate tariffs for the EQ-5D (among cancer patients), the ICECAP-O instrument and the CES carer instrument and is currently being used to replicate the ICECAP-O work in Australia using a large internet panel.[6, 31] All of these valuation exercises used a case 2 best-worst scaling (BWS) study.[32] This type of BWS is called the 'profile' case (although in early literature was called the 'attribute' case). Case 2 is attractive because it is potentially an easier choice task: respondents are presented with states one at a time and make choices *within* states, not between states. However, this leads to the same problem encountered in the smaller DCEs discussed above: length of life is not varied so the quality/quantity trade-off cannot be estimated. Thus similar second best re-anchoring solutions to those discussed already are required. It should be noted that vulnerable populations (such as children and older people) may find a theoretically correct DCE too difficult and that second best estimates are the only ones available. This is an empirical issue.

Researchers considering using case 2 BWS to value quality of life instruments should be aware that they may obtain different estimates to those from a traditional multi-profile choice task (even allowing for differences in variance scale factor). To explore this it is necessary first to consider the context of a choice task. Mathematical psychologists work within a paradigm that separates the importance of an attribute per se – which might vary depending on the context of the choice task –

from the level scale of an attribute level – which should be fixed in value not matter what the context of the choice.[30] The relationship between these is multiplicative, but the separation of the two is not part of traditional economic theory. The relative weights (for one context, relative to another) might be estimable if respondents answer a study twice, with context changing between the two: an example might be choice of airline, first when considered for pleasure, then for business. Conducting a case 2 BWS study (as opposed to a traditional multi-profile task) might change the context in a way that is inconsistent with the decision-making context in which the tariffs are meant to be used. This becomes clearer by considering what might happen had the ICECAP instrument (with five attributes of general quality of life):

(1) been valued using a pairwise DCE;

(2) included a sixth attribute, "political system", with two levels, "dictatorship" and "democracy".

It is possible that whilst "dictatorship" might always be picked as worst in the case 2 study (implying it is infinitely bad in a random utility model),[19] it might appear to be little worse than "democracy" in the pairwise traditional DCE. This is because respondents might focus more on 'obtaining the best individual quality of life they can' in their choices, with the result that that attribute rarely/never influences their choices. The small (zero) attribute weight for political system cancels the large difference in the level scales. The different choice task has changed the perspective of the respondent – (s)he no longer is identifying aspects of a given life (with certainty) that are 'bad' (or, more to the point, are so bad as to make it potentially unacceptable) but is being asked to make an inherently difficult trade-off: how much individual quality of life is (s)he willing to give up to obtain democracy? Some respondents might question the realism of such a choice (believing that no one individual can change the system) and essentially ignore the political system attribute. The need to consider the incentive compatibility of the choice has been discussed in the environmental economics literature;[33] whilst this example is highly unrealistic the possibility of a change in context being

induced by changing the nature of the choice task should be considered by designers of QALY valuation exercises.

Type 3 BWS represents an extension of a traditional 'choose one' DCE:[23] choice sets must be of size three or more, and the respondent is asked to choose the best profile/scenario (as in a traditional DCE) but also the worst profile/scenario. Therefore it presents some of the same practical problems that have discussed with respect to ranking studies and DCEs.

7. Design issues

The field of statistical design in choice experiments is moving forward rapidly. Principles of optimum design are described in detail elsewhere;[28] this section will describe some issues that are pertinent to QALY valuation exercises, particularly in the light of the variance scale issue. In the same way that the assumption that the MIXL model can adequately model any pattern of heterogeneity has proven misguided, it would be unwise for the researcher to assume that better G-MNL models can deal with all patterns of variance heterogeneity. *Ceteris paribus*, if (s)he can avoid a design that systematically induces large random utility terms in some choice sets and small ones in others, it would be wise to do so: simpler models (such as MNL) can then be estimated. For example, consider three EQ-5D states, 11111 (full health), 33333 (the 'pits' state) and 22222 (moderate problems on all five attributes). 11111 and 33333 are described more objectively than 22222 (which is characterised by more subjective amounts like 'some'); choices involving 22222 are likely to induce a larger random utility term than those involving 11111 or 33333. As a result, within the class of relatively poor health states respondents will be more certain as to what life is like in states with a larger number of attributes at level '3' than those with a smaller number; a smaller random utility term will increase (in absolute magnitude) the estimated decrements associated with level '3'

compared to level '2', *independently of and in addition to the actual (mean) decrement in utility*. Further examples of possible variation in the scale factor are provided in the context of the death state.[19] A potential solution is essentially Bayesian: by utilising prior information about the likely mean utilities of health states, the researcher can construct choice sets that attempt to be 'equally easy' (or 'equally difficult'!)

8. The research frontier

The field of discrete choice based tasks is going through a period of major methodological advance. The considerable difficulties in designing and analysing DCEs (and similar studies) mean that they are (or should be) increasingly considered a field in their own right, as opposed to 'just another' method of preference elicitation. There are some issues on the research frontier to which designers of CUA should pay particular attention.

8.1. Subgroup tariffs and individual level models

Designers of QALY valuation exercises should not ignore heterogeneity on the grounds that 'only' average population preferences are required. It is well known that estimates from discrete choice tasks cannot be properly aggregated until variance heterogeneity is adjusted for.[10, 34] Thus, the question is, "what are the implications of this need to properly model mean and variance heterogeneity?" This paper makes no recommendations about whether policy makers *should* use any such information: that is a normative issue. Rather, the aim is to describe new techniques to characterise heterogeneity and to outline (briefly) how the results might be useful to researchers and policy makers.

Recently variance scale adjusted latent class models have been implemented in the Latent Gold 4.5 software package. These models assign respondents to a variance class, as well as a mean class. The treatment of variance heterogeneity is somewhat simplistic: all a person's utilities are simply scaled up or down by a single variance scale factor (rather than allowing a respondent's consistency to vary by attribute). However, this has the statistical advantage of parsimony (for example two variance classes means only one additional parameter to estimate) and the conceptual advantage that it is consistent with a model in which some people are simply better at DCEs than others. Information on variance heterogeneity, in particular what sociodemographic variables are associated with it, is critical to future QALY valuation exercises: researchers should oversample those respondents likely to exhibit larger variances (lower choice consistency) to maintain precision and to guard against the possibility of identifying spurious mean heterogeneity. Scale-adjusted latent class models are also useful to policy makers since they are consistent with a policy question "*Do* the population average values conceal real – i.e. mean – subgroup heterogeneity?" If not, then the issue of whether or not to acknowledge heterogeneity becomes moot.

Recently, best-worst scaling has been used to estimate individual level utility functions.[23] Use of such models in CUA can avoid the problem of mis-characterising mean/variance heterogeneity entirely: the empirical distribution of true preference heterogeneity is available and no modelling is required. This has two benefits specific to QALY valuation. First, in the absence of longitudinal data, it offers a second-best opportunity to model any respondent adaptation to impaired states. If differences in preferences for particular impaired states are associated with experience of living in that state, then such differences can be quantified and, if found to have adverse equity implications for population values, can be adjusted for. The second benefit is an extension of the first: individual level models allow the researcher to test whether experience of states has *any* effect upon preferences, the critique of those who prefer the experienced utility paradigm to the decision utility

one.[35, 36] The key advantage underpinning these advantages is that differences in preferences allow more deliberation and debate, a policy aim of Sen,[37] which can address current issues such as whether “a QALY is a QALY is a QALY”.

8.2. Valuing whole life profiles and risk

Given non linearity in the utility of life duration,[38] health economists are interested in valuing ‘whole lives’ rather than just a period of impaired health, followed by death (or, even less realistically, a period of perfect health followed by immediate death). These attempts are welcome and offer the ability to better understand how the utility of life duration varies over the life cycle, amongst other issues. However, valuing these complete lives in a discrete choice framework presents considerable difficulties. The potential for variance heterogeneity is even greater: it is not unreasonable to assume that one’s preferences about the far future exhibit less consistency than those about the near future. Correctly modelling this will be essential if parameters such as the respondent’s discount rate are to be correctly estimated and modelled.

The empirical difficulties in dealing with risk via the standard gamble have been mentioned. The difficulties in doing so within a discrete choice framework are no less pertinent and the author is part of a team that have applied for funding to conduct a large study utilising randomised trials to inform this issue.

8.3. Interactions

Many DCEs in health care use Orthogonal Main Effects Plans (OMEPs). It is well established in other fields that the assumptions required for the estimates from OMEPs to be unbiased often do not

hold.[25] Therefore there is an urgent need for researchers to conduct larger studies capable of estimating (at least) two-way interactions. Indeed the ‘N3’ term in the EQ-5D regression models of various countries suggests that interactions exist for that instrument at least.

9. Conclusion

Scientists should try to use those models that require fewer assumptions than those that require more. Therefore, recent interest in the use of discrete-choice based tasks that assume only ordinality rather than cardinality to estimate QALY values is welcome. This paper has attempted to improve awareness among the health economics community of some serious issues that are endemic to these ranking, discrete choice and best-worst scaling studies. It has referenced papers in the environmental economics, econometrics, marketing and psychology literature that are of critical importance if health economists are to understand the limitations of these studies. Indeed a key marketing paper shows that the MIXL regression model commonly used by health economists is flawed in real and simulated data.

Crucially, health economists should recognise that interpretation of logistic regression output is fundamentally different in DC-based tasks from that of a disease model: variance heterogeneity has implications for bias and not just efficiency of the (latent mean) estimates. This important point was recently communicated excellently to the sociology community.[39] Health economists should take note.

In terms of positive recommendations, the author would suggest three to aid research teams who wish to estimate QALY tariffs from DC-based tasks:

(1) Ensure all respondents answer a common main effects design but block the set of profiles that are capable of estimating interactions; randomise respondents to blocks.

(2) Ask enough questions (whether by best-worst techniques or by presenting more choice sets) to give better insights into both mean and variance heterogeneity. Ideally, attempt to estimate individual level decision rules.

(3) Use statistical models that are appropriate to the psychological models being used by respondents; do not take a 'one size fits all' approach by using one statistical model and try, wherever possible, to identify which psychological model is being used through the use of structured questioning.

Rapid changes in what constitutes 'good practice' in DCEs can be frustrating. However, designers of QALY valuation exercises have the opportunity to ensure the dissemination of such standards quickly and widely: large studies that elicit a lot of information from each respondent should be less vulnerable to problems of inference and analysis. Furthermore, such studies will naturally have the power to test a number of assumptions made in current QALY models, and should help inform a number of current debates, including whether experienced of decision utility should be the method by which preferences are elicited.

References

1. Arnesen, T and Trommald, M. (2005). Are QALYs based on time trade-off comparable? - A systematic review of TTO methodologies. *Health Economics, 14,* 39-53.
2. Richardson, J. (1994). Cost utility analysis: what should be measured?. *Social Science & Medicine, 39(1),* 7-21.
3. Luce R. D. (1959). *Individual choice behavior*. New York: John Wiley & Sons.
4. Williams A. (1995). The measurement and valuation of health: a chronicle. York: University of York; *Report No. 136*.
5. Ryan, M., Netten, A., Skatun, D., and Smith, P. (2006). Using discrete choice experiments to estimates a preference-based measure of outcome - An application to social care for older people. *Journal of Health Economics*.
6. Coast, J., Flynn T. N., Natarajan, L., Sproston, K., Lewis, J., and Louviere J. J. (2008). Valuing the ICECAP capability index for older people. *Social Science & Medicine, 67,* 874-82.
7. Ben-Akiva, M., Lerman, S. R. (1985). *Discrete choice analysis: theory and application to travel demand*. Cambridge, MA: MIT Press.
8. Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34,* 273-86.
9. McFadden, D. and Zarembka, P. (1974). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*. New York: Academic Press, p. 105-42.
10. Hensher, D. A., Louviere, J. J., and Swait, J. Combining sources of preference data. *Journal of Econometrics,89,* 197-221.
11. Louviere J. J. (2006). What you don't know might hurt you: some unresolved issues in the design and analysis of discrete choice experiments. *Environmental and Resource Economics,34,* 173-88.
12. McFadden, D. and Train, K. (2000). Mixed MNL Models For Discrete Response. *Journal of Applied Econometrics, 15(15),* 447-70.
13. Fiebig, D., Keane, M., Louviere J. J., and Wasi N. (In press). The Generalized Multinomial Logit Model. *Marketing Science*.Working paper available at: http://www.censoc.uts.edu.au/researchoutput/wp09002.pdf
14. Luce, R. D., Suppes, P., Bush, R. R., and Galanter, E. (1965). Preference utility and subjective probability. *Handbook of mathematical psychology,* volume III. New York: Wiley, p. 249-410.
15. Marley A. A. J. (1968). Some probabilistic models of simple choice and ranking. *Journal of Mathematical Psychology,5,* 23.
16. Beggs S., Cardell S., and Hausman J. (1981). Assessing the potential demand for electric cars. *Journal of Econometrics, 16,*19.
17. Salomon J. A. (2003). Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. *Population Health Metrics, 1(12).*
18. McCabe C., Brazier J. E., Gilks, P., Tsuchiya, A., Roberts, J., and O'Hagan, A. (2006). Using rank data to estimate health state utility models. *Journal of Health Economics, 25,* 418-31.
19. Flynn, T. N., Louviere, J. J., Marley, A. A. J., Coast, J., and Peters, T. J. (2008). Rescaling quality of life values from discrete choice experiments for use as QALYs: a cautionary tale. *Population Health Metrics, 6,*1-6.
20. Hausman, J. A., Ruud, P. A. (1987). Specifying and Testing Economic Models for Rank-Ordered Data. *Journal of Econometrics, 34,* 83-104.
21. Ben-Akiva, M., Morikawa, T., and Shiroishi, F. (1991). Analysis of the Reliability of Preference Ranking Data. *Journal of Business Research, 23,* 253-68.

22. Finn, A. and Louviere, J. J. (1992). Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety. *Journal of Public Policy & Marketing, 11(1),* 12-25.
23. Louviere, J. J., Street, D. J., Burgess, L., Wasi, N., Islam, T., and Marley, A. A. J. (2008). Modelling the choices of single individuals by combining efficient choice experiment designs with extra preference information. *Journal of Choice Modelling, 1(1),* 128-63.
24. Helson, H. (1964). *Adaptation-Level Theory*. New York: Harper & Row.
25. Louviere, J. J., Hensher, D. A., Swait, J. (2000). *Stated choice methods: analysis and application*. Cambridge: Cambridge University Press.
26. Lancsar, E., Louviere, J. J. (2008). Conducting discrete choice experiments to inform healthcare decision making. A user's guide. *Pharmacoeconomics*. *26(8),* 661-77.
27. Bleichrodt H. (2002). A new explanation for the differences between time trade-off utilities and standard gamble utilities. *Health Economics, 11,* 447-56.
28. Street, D. J., Burgess, L. (2007). *The Construction of Optimal Stated Choice Experiments: Theory and Methods*. John Wiley & Sons Inc.
29. Marley, A. A. J., Louviere, J. J. (2005). Some probabilistic models of Best, Worst, and Best-Worst choices. *Journal of Mathematical Psychology, 49,* 464-80.
30. Marley, A. A. J., Flynn, T. N, and Louviere, J. J. (2008). Probabilistic Models of Set-Dependent and Attribute-Level Best-Worst Choice. *Journal of Mathematical Psychology, 52,* 281-96.
31. Szeinbach, S. L., Barnes, J. H., McGhan, W. F., Murawski, M. M., and Corey, R. (1999). Using conjoint analysis to evaluate health state preferences. *Drug Information Journal, 33,* 849-58.
32. Flynn, T. N., Louviere, J. J., Peters, T. J., Coast, J. (2007). Best-Worst Scaling: What it can do for health care research and how to do it. *Journal of Health Economics, 26(1),* 171-89.
33. Carson, R. T. and Groves, T. (2007). Incentive and informational properties of preference questions. *Environmental and Resource Economics, 37,* 20.
34. Swait, J. and Louviere, J. J. (1993). The role of the scale parameter in the estimation and comparison of multinomial logit models. *Journal of Marketing Research, 30,* 305-14.
35. Dolan, P. (2008). Developing methods that really do value the 'Q' in the QALY. *Health Economics, Policy and Law, 3,* 69-77.
36. Dolan, P. and Kahneman, D. Interpretations of utility and their implications for the valuation of health. *Economic Journal, 118,* 215-34.
37. Sen, A. (2005). Human rights and capabilities. *Journal of Human Development, 6,* 16.
38. Attema, A. E. and Brouwer, W. B. F. (2009). The correction of TTO-scores for utility curvature using a risk-free utility elicitation method. *Journal of Health Economics, 28,* 10.
39. Mood, C. (In press). Logistic Regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*.