Choice experiments in health: the good, the bad, the ugly and toward a brighter future

JORDAN J. LOUVIERE*
Professor of Marketing, Executive Director, Centre for the Study of Choice, University of Technology, Sydney, Australia

EMILY LANCSAR
Lecturer in Economics, Business School and Institute of Health and Society, Newcastle University, UK

Abstract: Compared to many applied areas of economics, health economics has a strong tradition in eliciting and using stated preferences (SP) in policy analysis. Discrete choice experiments (DCEs) are one SP method increasingly used in this area. Literature on DCEs in health and more generally has grown rapidly since the mid-1990s. Applications of DCEs in health have come a long way, but to date few have been ‘best practice’, in part because ‘best practice’ has been somewhat of a moving target. The purpose of this paper is to briefly survey the history of DCEs and the state of current knowledge, identify and discuss knowledge gaps, and suggest potentially fruitful areas for future research to fill such gaps with the aim of moving the application of DCEs in health economics closer to best practice.

Introduction

Compared with many applied areas of economics, health economics has a strong tradition in stated preference (SP) elicitation and use in policy analysis. Examples include SP data elicited using time tradeoff, person tradeoff, standard gamble and contingent valuation methods (Drummond et al., 2005). Use of SP data partly stems from necessity because data sources traditionally used by economists, namely revealed preferences (RP) or observations of choices made in real markets, are limited in health due to lack of traditionally functioning markets. Even in sectors where RP data are available, there is growing recognition that SP data can substitute for and complement RP data.

Discrete choice experiments (DCEs) are one SP method increasingly used in applied economics to address key policy issues. Typically, DCEs involve

*Corresponding author: Jordan J. Louviere, Professor of Marketing, Executive Director, Centre for the Study of Choice, University of Technology, Sydney, PO BOX 123, Broadway, NSW 2007, Australia. Email: Jordan.Louviere@uts.edu.au
participants choosing among hypothetical (occasionally real) options presented in surveys where the options are described by characteristics or attributes. For example, diagnostic tests can be described by the attributes’ test accuracy, time to receive results and price. Each attribute takes on a range of levels; for example, time for results can be described by hours, days or weeks. Attribute levels are combined to produce hypothetical options (profiles) and assigned to choice sets using experimental designs. Participant choices of price- and quality-differentiated options are analysed with discrete choice models in a random utility theory (RUT) framework to estimate the preferences implied by a particular indirect utility function (IUF).

Preference-model estimates can be used for policy analyses, such as investigating the relative impacts of health-care product/programme attributes (or indeed a health state), or calculating rates at which people are prepared to trade off less of one attribute to obtain more of another (e.g. accept more waiting time to obtain more accurate diagnostic test results). The latter is particularly relevant in resource-constrained environments like health. SP methods are also used to predict likely uptake of new health-care products/programmes (e.g. new cancer screening programmes), and to calculate measures of welfare gain, or willingness-to-pay (WTP), for individual attributes and/or entire health-care products/programmes.

Indeed, DCEs have been used to address each of these issues in health economics to various degrees, with varying success. Despite growing popularity and application in health settings [see Ryan and Gerard (2003) and Viney et al. (2002) for reviews], knowledge gaps remain that can be better appreciated if placed in a historical context to understand how SP methods have been used, and the current research frontier. Thus, the objectives of this paper are to: (a) review DCE history in general, and its evolution in health; (b) briefly review the current state of knowledge, identify and discuss knowledge gaps; and (c) suggest potentially fruitful areas of future research. The contribution of the paper is to try to move DCEs in health closer to best practice in the broader context of DCE applications more generally.

**Brief history**

Discrete choice experiments have a long history in several disciplines and it is useful to review their development to better understand and appreciate their applications. Thurstone (1927) pioneered RUT and the method of paired comparisons. The latter was a workhorse SP elicitation experiment/task for decades, and continues to be used despite being dominated by DCEs for most applications. Indeed, advances in paired comparisons continue in psychology and statistics; an excellent summary of some of these is provided by David (1988).
In the 1960s, work began on ordered relations on sets, culminating in the theory of conjoint measurement (Luce and Suppes, 1965). The aim of this work was to develop ways to represent what economists term ‘indirect utility functions’ and psychologists term ‘decision rules’ algebraically. A more specific aim was to represent a ranking of a set of multi-attribute profiles with specific algebraic forms, like strictly additive, multiplicative, dual-distributive, and other such models. Philosophically, it was widely believed at the time that humans could reliably and validly only ‘rank’ a set of options; so, any monotonic transformation of ranking responses consistent with a hypothesised model was acceptable. Unfortunately, allowing ‘any’ monotone transformation to satisfy specific algebraic forms raises issues about whether humans differ in such transformations and whether such differences generalise. To our knowledge, such issues remain unanswered.

In parallel, two other research streams impacted modern thinking in DCEs, namely Social Judgment Theory (Brunswick, 1955) and Information Integration Theory (Anderson, 1981). Both paradigms still have adherents and can be viewed as forms of what has come to be called ‘conjoint analysis’. The latter term was coined by Green and Rao (1971), and came to be applied to any SP elicitation method that uses some way to combine attribute levels into profiles, and asks samples of people to give some type of preference response to the profiles implemented in surveys. Unfortunately, this description is so general that it renders the term ‘conjoint analysis’ essentially meaningless as a specific descriptor of any particular SP elicitation approach.

While it can be argued that a behavioural theory underlies Social Judgment Theory and Information Integration Theory, the original motivation for ‘conjoint analysis’ methods was the purely axiomatic theory of conjoint measurement, which is a theory about the behaviour of numbers, specifically orders on sets, rather than the behaviour of humans. If the rankings provided by humans satisfy the underlying axioms of conjoint measurement theory, they can be represented ‘as if’ humans added, multiplied, and so forth, utilities associated with each attribute level to assign a ranking to each combination of attribute levels (profile). Conjoint measurement was embraced by academic and practicing marketing researchers, but has since morphed into a set of ad hoc techniques that now (at best) rely on statistical theory to justify applications; virtually any pretence of relying on behavioural theory was abandoned by the early 1980s.

Also, in the 1960s, work continued in the Thurstone tradition of modelling choice behaviour, with Luce (1959) and Restle (1961) providing well-known work that extended Thurstone’s paired-comparison ideas to multiple choices. However, the latter work was not in the random-utility paradigm, but in a fixed-utility paradigm. The former views preferences as fixed, but choices as variable; whereas the latter views preferences as fixed but only partially observable. Both theories lead to forms of probabilistic discrete choice models that can...
be used to understand, explain and predict choices made by individuals and groups. Choice-theoretic models provide behaviourally justifiable ways to model preferences that contrast with the non-behavioural, statistical underpinnings of conjoint analysis. They also have an associated error theory and can be applied to real (revealed) or stated (hypothetical) choices.

With this background, we can better understand the development of multi-attribute DCEs pioneered by Louviere and Woodworth (1983). This DCE approach grew out of the need to solve practical problems (primarily in Australia in the late 1970s) that could not be solved with conjoint analysis. Louviere and Woodworth (1983) pioneered the use of experimental design theory to construct choice sets of profiles that participants evaluate and choose among (rather than rank or rate, as in conjoint analysis), and they noted that participant choices from DCEs were consistent with conditional logit models (CLMs; McFadden, 1973). Prior to this time, paired comparisons and ad hoc sets of pairs, triples, quadruples, and so on, were used to study multiple choices, particularly in psychology and statistics (Kendall and Smith, 1940; Bradley and Terry, 1952; Wells, 1991). For the first time, Louviere and Woodworth (1983) provided a systemic way to construct choice sets so that the choice data were consistent with particular types of statistical choice models, and could be used to simulate a wide array of real markets.

It is worth noting that shortly after the Louviere and Woodworth paper appeared, the DCE approach was labelled as ‘just another form of conjoint analysis’ (which is incorrect), with the term ‘choice-based conjoint analysis’ used to describe what we now call ‘DCEs’. The latter is responsible for much confusion, that continues to this day. Lancsar and Louviere (2008b) compare DCEs and conjoint analysis, noting key differences in these methods and their implications for preferences and welfare analysis.

Little additional progress was made on ways to design DCEs until the mid-1990s; a notable exception being availability designs pioneered by Anderson and Wiley (1992). Papers on design efficiency began to appear in the mid-1990s, with a widely-cited working paper by Bunch et al. (1996), sparking interest in ‘shifted designs’, which were later shown to be associated with optimal design-construction methods (Street and Burgess, 2007). Huber and Zwerina (1996) proposed what they viewed as desirable design criteria for DCEs, one of which is ‘utility balance’ that involves designing choice sets such that choice options are as similar in utility as possible. It is surprising that this idea has persisted because if this criterion is satisfied, one obtains no reliable statistical information to estimate models.

The DCE method was introduced to health economics by Propper (1990) and advocated by Ryan and colleagues in the late 1990s (e.g. Ryan and Hughes, 1997; Ryan, 1999). The initial role of DCEs in health was as a way to study preferences and values associated with process characteristics and non-health outcomes (Ryan, 1999), with less attention paid to health outcomes. Indeed, a
key reason for applying DCEs at the time was to ‘go beyond the Quality Adjusted Life Years (QALY) framework’ that was and remains the dominant form of evaluation in health economics. While there was some methodological work, the early focus was on applications, primarily to elicit patient preferences and tradeoffs for attributes describing a good or service of interest in various clinical settings (e.g. Ryan and Hughes, 1997; Ryan, 1999; Johnson et al., 2000). Other health applications included eliciting preferences for time (e.g. van der Pol and Cairns, 2001), health service configurations (e.g. Jan et al., 2000; Morgan et al., 2000; Bech, 2003), health insurance packages (e.g. Chakroborty et al., 1994) and GP remuneration packages (e.g. Scott, 2001), to name a few.

The last decade has seen more methodological work and recognition that DCEs can contribute more than just information on preferences and values associated with process characteristics. A broader focus is also evident, such as predicting uptake of new policies or programmes (e.g. Hall et al., 2002) and valuing health outcomes and tradeoffs (e.g. Gyrd-Hansen, 2004; Lancsar et al., 2007b). Interestingly, and perhaps uniquely, DCEs can potentially contribute both to cost–benefit analysis (CBA) and cost-effectiveness analysis (CEA) (Lancsar and Louviere, 2008a), a topic we return to in the future research section.

Useful reviews of the DCE literature in health can be found in Viney et al. (2002), Ryan and Gerard (2003) and Ryan et al. (2008). Lancsar and Louviere (2008a) provide a recent ‘how to guide’ that reviews the DCE approach and offers advice on factors to consider in undertaking and assessing the quality of a DCE, including a detailed checklist.

**Current state of knowledge**

We discuss the current state of play of DCEs in general, and more specifically in health, focusing on two (arguably, the most) important components of DCEs – experimental design and estimation of associated choice models.

**Experimental design**

Street and Burgess (Street and Burgess, 2004; Burgess and Street, 2005; Street et al., 2005) contributed a systematic focus to optimal design of DCEs with optimal-design theory for generic forced-choice experiments and CLMs. The statistical efficiency properties also are known for certain DCE designs that vary price (e.g. Kanninen, 2002). Despite progress, optimal DCE design is in its infancy, with many unresolved problems noted later. Our knowledge of optimal designs for problems involving strictly additive, generic IUFs is relatively robust; but little is known about efficiency properties for designs for other IUFs or associated-choice models. Identification properties of the so-called L
designs for alternative-specific and availability problems also are reasonably well-understood (Louviere et al., 2000), but as yet, there are no results on efficiency for models other than CLMs.

Experimental design is a relatively new methodological area for most economists that is under-explored in health, with some exceptions (Carlsson and Martinsson, 2003; Maddala et al., 2003; Viney et al., 2005). Many earlier health studies used small designs to construct DCEs (Ryan and Gerard, 2003), most of which relied on what are known as ‘orthogonal main-effects plans’ (OMEPs) that are fractional factorial designs that only allow estimation of attribute main effects. To use an OMEP to construct a DCE, one must assume that ‘all’ attribute interactions are insignificant, which is unlikely to be satisfied; and if any omitted interactions are significant, parameter estimates are biased. Thus, OMEPs restrict estimation to strictly additive IUFs and usually result in high ratios of parameters to observations that should suggest caution. Consequently, health researchers are starting to use larger designs (e.g. King et al., 2007), a trend that should continue. In fact, one can argue that the largest design possible should be used for a given time and resource constraints. At a minimum, designs that allow independent estimation of all main effects and two-way interactions ensure that even if two-way interactions are significant, they are independent of main effects, minimising bias if interactions are omitted. Indeed, one can often use full factorial designs (all attribute-level combinations) by creating mutually exclusive and exhaustive blocks/versions, and randomly assigning participants to different versions. Thus, OMEPs should be used only if it is impossible to use larger designs, or if one must compare participants’ responses to a common design.

It is important to note that experimental designs for DCEs involve combining attribute levels to form profiles (sometimes called a ‘starting design’) and optimally assigning them to choice sets. Profiles in health economics have been designed in several ways: sourcing designs from catalogues, ‘hand’ construction of designs, using software (e.g. SAS, SPEED and SPSS) and generating choice sets in various ad hoc ways. Below, we consider some ad hoc but commonly used ways to assign profiles to choice sets, and discuss the implied statistical properties in more detail.

It is also worth noting that Burgess et al. (2006) and Street et al. (2008) reviewed work in health economics and elsewhere, and found many examples of statistically inefficient designs and/or designs with identification problems. Street and Burgess (2007) also evaluated design-generation methods and noted that several of them often give inefficient designs. They also found that design methods available in SAS often produce efficient designs, but the effects are not fully independent, resulting in some attribute effects with shared covariances. Thus, commercially available design-generation software should be viewed as a tool, not a ‘solution’. One solves a particular design problem by understanding how to use such tools, and the implications and consequences of their underlying assumptions.
Now we discuss some examples of poor design strategies previously used in health to assign profiles to choice sets, and why they should be avoided. A common strategy was to create or obtain a starting design with a number of profiles and randomly pair one profile with all of the others. For example, one might create eight profiles and randomly pair one of them with the others to make seven pairs. This strategy can lead to identification problems, and Louviere et al. (2008a) show that it can produce non-orthogonal designs whereby one or more main effects are confounded with other main effects. In such cases, attribute effects of interest may not be separately identified, which can also pose problems in identifying and estimating interactions.

Another approach used to make choice sets was to randomly pair profiles from a starting design. Continuing the eight-profiles example, this involves randomly pairing them with each other to make choice sets. If all pairs are used (28), all effects supported by a starting design are identified and can be estimated; problems arise when researchers use some ad hoc number less than all pairs, as this can lead to identification issues. An alternative design strategy is to construct all possible pairs from a starting design and randomly select from them in such a way that all the pairs are observed across the DCE (preferably more than once each). Louviere et al. (2008b) illustrate this approach with a full factorial for four two-level attributes (16 profiles) that has 120 possible pairs \( \frac{n(n-1)}{2} \). There are many possible ways to choose from the 120 pairs, which they simplify by restricting the possible choices to eight pairs each; they took 1000 random draws of eight without replacement and calculated the efficiency of each set of eight pairs. They found that many sets of pairs are very inefficient; hence, this strategy can give a good design by chance, but one is very likely to obtain poor designs.

OMEPs are constructed in such a way that design columns used to represent the main effects exactly equal one or more interactions of main effects. Many researchers seem unaware that the effects that one estimates from such design will represent a main effect of interest if, and only if, all omitted interactions are not significant. Thus, model parameters estimated from such designs may represent (a) a main effect of interest, or (b) one or more interactions and/or (c) a linear combination of the main effect of interest and unobserved interaction(s). Thus, if one uses these approaches to design DCEs, any problems are ‘self-inflicted’ in so far as a researcher ‘designs’ poor properties into a DCE. Such practices are surprising in light of the fact that a key benefit of DCEs is they can (and should) be used to generate choice data with good instead of bad statistical properties.

**Estimation of choice models**

Largely unrelated to progress in experimental design, major developments have occurred in types of choice models that can be estimated from choices in DCEs.
McFadden’s CLM continues to be applied despite results in the wider choice-modelling literature, indicating that the underlying assumptions are rarely satisfied. To wit, assumptions about systematic and error components can be relaxed in two general ways (a) more general error variance–covariance matrices, including nested logit, non-identity multinomial probit and heterogeneous error-variance models; and (b) allowing model parameter estimates (fixed effects in CLM) to vary randomly across sample populations according to continuous (e.g. normal) or discrete distributions, which leads to, respectively, the Mixed Logit model (MIXL) (Revelt and Train, 1998) or its Hierarchical Bayes (HB) counterpart (Allenby and Rossi, 1998; Greene et al., 2006) and finite mixture or latent class (LC) models (Kamakura and Russell, 1989). New ways to capture dynamics have also led to more sophisticated models of choice evolution over time (Keane and Wolpin, 1994; Erdem and Keane, 1996). CLMs have frequently been used to analyse DCE data in health, but other model forms, such as nested logit (e.g. Ryan et al.), random parameter probit (e.g. Johnson et al., 2000), MIXL (e.g. Hall et al., 2006; King et al., 2007; Lancsar et al., 2007b; Hole, 2008; Kjaer and Gyrd-Hansen, 2008), and recently LC (Hole, 2008), have also been applied.

Finally, recent developments in theory, methods and designs for DCEs consistent with Best–Worst (BW) Scaling (Marley and Louviere, 2005; Flynn et al., 2007; Marley et al., 2008) have begun to appear. Introduced by Finn and Louviere (1992), BW tasks are a type of DCE in which participants choose: (a) the best and worst (most and least preferred) options from sets of three or more options, (b) the best and worst attribute levels from single options or profile descriptions (McIntosh and Louviere, 2002; Flynn et al., 2007; Lancsar et al., 2007a) or (c) the best and worst options from sets of three or more multi-attribute options (Lancsar and Louviere, 2005, 2009). Designs for BW DCEs include balanced, incomplete block designs for case (a), full or fractional factorial designs for case (b) and traditional DCE designs for case (c).

**Open issues and research questions for DCEs**

Now we address what needs to be done, unresolved issues and potentially fruitful areas for ongoing research to address such issues.

**Pre-design**

There seems to be broad agreement that pre-design qualitative work is needed to identify attributes and levels and understand words/phrases that decision-makers use to describe the attributes and levels, and associated decision processes. An important and relatively neglected area of inquiry is the best way to go about doing this. It is common in health to identify attributes and levels with reviews of relevant literature, sometimes supported by ad hoc qualitative
research like focus groups or interviews with relevant populations and/or experts. The field needs theory, or at least a more systematic approach to qualitative research, for developing attributes and levels and choice contexts (e.g. Coast and Horrocks, 2007). Identification and selection of attributes and levels requires iterative testing and refining to ensure that they make sense and are interpreted in the same way by all DCE participants. Changes to (and more research on) this aspect of DCEs would be welcome.

Researchers often try to measure ‘how important’ each of the potentially many candidate attributes might be, before designing experiments. Yet, one sees little evaluation of whether attributes identified and prioritised in pre-design work predict real choices and/or whether attribute effects revealed in DCEs relate to a priori ‘importance’ measures; hence, these clearly are potentially fruitful avenues for further research.

Of further interest is how best to simulate actual decision process(es). For example, asking participants to choose one of a pair of options described by different attribute levels may be common practice in health economics, but may not closely simulate a real decision context. Forced choices from pairs, triples or larger sets without an option not to choose or stay with a status quo may be inconsistent with economic theory and may not be externally valid. Ryan and Skatun (2004) used nested logit to model choices of participants who preferred none of the hypothetical options offered in a DCE. There seems to be a clear need for comprehensive theory and/or standards-based-on-theory or accumulated evidence to guide researchers in attribute identification, assignment of levels, task instructions, task layout and formatting.

Applying DCEs to policy contexts requires a thorough understanding of the policy question to inform DCE design and interpretation of the associated preference model(s). This includes phrasing of the type of good (e.g. public or private goods) that includes consideration of appropriate attributes and levels, and an appropriate payment vehicle(s) if price is involved. In turn, this depends on the type(s) of good evaluated, consideration of whose preferences are relevant and choice of appropriate sample from the population, and participants’ likely level of understanding and familiarity with the topic.

**Design**

Despite significant advances in DCE optimal-design theory (e.g. Street and Burgess, 2007), design capabilities lag behind model developments, with optimal-design theory for more general models and IUFs undeveloped. Work is needed on optimal designs for alternative-specific or labelled choices, as well as DCEs with individual-specific status quo options.

The overriding DCE design objectives should be identification and efficiency. Identification determines which model effects can be estimated, and informs analysts about which forms of IUFs can be estimated conditional on a particular
form of choice model. Efficiency refers to the precision with which the effects that are identified can be estimated, and more efficient designs give more precise parameter estimates for a given sample size. Given our current knowledge about the consequences of violating maintained assumptions associated with designs discussed below, we recommend that one first focus on identification, and then on efficiency, because one may be able to improve efficiency by increasing sample size, but identification cannot be changed once a design is constructed.

There also seems to be a general confusion over design properties. Statisticians have provided an ‘alphabet soup’ of design criteria where each ‘letter’ indexes a statistical property to be optimised. A widely used criterion is ‘D-Optimality’, or optimising the determinant of the Fisher Information Matrix to minimise the generalised variance of the parameter estimates. Typically, this gives reasonably robust designs for most DCE applications, but there may be cases where one wants to optimise more specific criteria for one or more model parameters. To further confuse applied researchers, one can optimise design with classicist or Bayesian methods (Carlsson and Martinsson, 2003; Bliemer et al., 2008); more research is needed to identify general conditions where a priori knowledge of signs and/or true parameters will significantly improve design properties. It is also important for researchers to understand that ‘all’ current design construction methods rely on strong assumptions about true underlying processes for systematic and random utility terms. It is unclear if and when these assumptions are satisfied empirically. Indeed, it is disingenuous of design researchers to claim superiority of design-construction methods without also demonstrating that their maintained assumptions are empirically satisfied. Thus, research into statistical consequences of violating assumptions and/or design approaches that are as robust as possible, to departures from assumptions would be welcome, and should be a research priority.

Implausible attribute-level combinations also deserve research attention, and this issue may be more relevant to health than other fields. To wit, attributes and/or levels may be highly correlated and/or it may be implausible or impossible for certain levels to co-occur due to the science/technology in real markets. For example, combining the level ‘80 years of age’ for the attribute ‘age of onset of illness’ with the level ‘50 years’ for the attribute ‘increased life expectancy with treatment’, is implausible. Addressing such attribute and level combinations complicates design construction and can impact statistical properties of designs. Thus, we need research on the best ways to do this, especially as common practice of eliminating the nonsensical combinations may not always be a good idea. It is worth noting that it often will be the case that some combinations of levels are less ‘clear-cut’ than the above example; that is, it often will be the case that participants (not experts, clinicians, etc.) do not know levels are implausible and/or one can tell a logical/sensible story that makes them plausible. If pilot tests reveal that participants do not find them implausible, statistical design properties should receive priority; if participants recognise
the unrealistic profiles, one must find a sensible way to design the DCE to deal with it. More research is needed on the best way(s) to do this.

Discrete choice experiments may cause unintended interactions between participants and design aspects that may or may not be related to statistical design properties. For example, design-construction methods can produce more/less complex choice tasks, leading to issues of tradeoffs between statistical and respondent efficiency. For example, Street and Burgess’ (2007) designs typically result in all or most attribute levels differing among options in each choice set. In principle, this forces participants to evaluate all options on all attributes, increasing task complexity. Variability in participants’ choices can increase with task complexity, decreasing design efficiency even for optimally efficient designs. Moreover, there is evidence of systematic relationships between unobserved choice variability and aspects of designs (Louviere et al., 2008b). So, one must recognise that experimental designs, DCE tasks, DCE task instructions, layouts, formats, and so on, may all impact unobserved variability, decreasing statistical efficiency. Thus, research is needed into the tradeoff between better statistical efficiency and more choice variability. In some cases, it may be better to use less statistically efficient designs to make simpler choice tasks that lower unobserved choice variability; however, more research is needed into cases where this is appropriate.

Experimental design is perhaps the most important aspect of DCEs as this determines what model(s) can be estimated with what levels of precision. We frequently find that key aspects of experimental designs are not disclosed or are insufficiently disclosed in academic papers. It is important for researchers to recognise that it is impossible for reviewers to fully understand and assess the contribution (or lack thereof) of any given study without this disclosure.

**Implementation**

‘Implementation’ of DCEs refers to all activities involved in pre-design and design stages to final DCE surveys. We discuss sampling, survey methods and data collection. Little work has been done on sampling in DCEs other than where all participants answer the same choice sets, which allows sample size to be calculated (Louviere et al., 2000). If a design is blocked into versions, and sub-samples are randomly allocated to versions, little work is available to guide researchers. This is partly due to the non-linear nature of choice models, as one needs to know the true parameters in advance to estimate the sample size(s) to achieve particular confidence levels. It is also partly due to different sample size implications for different choice models and/or IUFs, even if true parameters are known in advance. So, sampling theory for the most general choice model and indirect utility functional forms is needed. However, this is likely to be a difficult problem.
Many health DCEs were administered as mail surveys and/or in-person interviews (Ryan and Gerard, 2003); some recent DCEs were completed online (e.g. Cairns and van der Pol, 2004). Literature on mail and in-person interview methods is vast, and we do not discuss it here. Instead, we note that administering DCEs online has several advantages, including minimising human error in coding, data entry and data management due to process automation, as well as faster data collection. Random samples from reliable web-panel providers differ little from population statistics, and survey questions that mimic census questions allow re-weighting to account for sample and population differences. That said, more research is needed on the reliability and accuracy of online surveys relative to other DCE health surveys.

Analysis

As earlier noted, some unrealistic assumptions of CLMs can be relaxed with new models like MIXL, HB and LC. More research is needed to compare such models empirically [e.g. Hole (2008) compared MIXL and LC], and determine if and when their associated assumptions are satisfied. For example, evidence is growing that error variances may not be constant within or between individuals (Louviere and Engle, 2006; Adamowicz et al., 2008; Fiebig et al., 2009; Louviere and Meyer, in press). If constant error-variance assumptions are violated, taste parameter distributions are confounded with unobserved variability distributions (Louviere, 2004), posing serious bias issues in using estimates of price or other welfare numeraire effects in policy analyses, because it can lead to preference parameter estimates for individuals with low/high unobserved variability that are significantly higher/lower than the ‘true’ estimates. WTP estimates may be biased if variability in choice within people is not separated from taste parameter variability across people. Yet, many researchers in health and applied economics assume that the only source of unobserved choice variability is preference heterogeneity. Hence, recognising that variability in tastes is only one of the many potential sources of variability in choice is needed, as is research into a theory of what underlies differences in choice variability in different DCE contexts.

Work to relax constant error-variance assumptions is proceeding on four fronts: (a) allowing more complexity in taste parameter variance–covariance matrices; (b) specifying models that separate within- and between-person variability; (c) specifying heterogeneous error-variance models and (d) estimating models for single persons. Work on option (a) is less attractive for several reasons, such as requiring more ‘latent’ parameters to be estimated, and strong assumptions on evolution (or not) of taste parameters over time and space. Fiebig et al. (2009) specified and applied a more general variance-components model called Generalised Multinomial Logit (G-MNL) that captures different types of within- and between-person variability. They applied G-MNL to several health datasets and showed that within-person components can be large.
DeShazo and Fermo (2002), Islam et al. (2007) and Swait and Adamowicz (2001) specified heterogeneous error-variance models, and Islam et al. (2007) showed that such models outperform MIXL in cross-sample prediction tests even when MIXL fits better in-sample. More work on these models is needed to determine if and when they are good approximations to what is likely to be a complex underlying process. Louviere et al. (2004, 2008b) developed ways to estimate models for single individuals, and recent work suggests that this provides better in-sample and out-of-sample fits than MIXL or HB (Louviere et al., 2009); more work is needed to determine the extent of possible applications and properties. Work has commenced in this area in health economics, as Lancsar and Louviere (2009) estimated individual-level choice models in a health setting and in doing so used a new sequential BW multinomial logit-estimation method.

Discrete choice experiments pose particular problems for modelling choices because they are repeated measures experiments; hence, choices should be correlated over successive choice sets. If one has more than one replication of a DCE (or can expand data as in Louviere et al. (2008b)), one may be able to estimate MIXL or HB models for single persons. In this case the key variance component is the variability in a person’s choices relative to the taste parameters, which in turn allows for inferences about whether a person’s choices are sufficiently consistent to give reliable mean estimates. New developments in variance component models are likely, and research into such models would be welcome.

A related stream of research has focused on asking BW preference questions as a way to obtain more preference data for a person, facilitating estimation of BW choice models for single persons (Louviere et al., 2008b; Lancsar and Louviere, 2009). Applications of this approach are likely to increase, and research into the limits to which one can go in terms of attributes, levels, choice sets and extra BW questions without compromising reliability and validity, would be welcome. Other BW methods can be used to measure all attribute levels on common scales, permitting inter-dimensional utility comparisons without a numeraire (McIntosh and Louviere, 2002; Flynn et al., 2007; Lancsar et al., 2007a). Research comparing these BW methods with welfare-theoretic measures, to understand when and why they might differ, would be welcome.

Validity

Research on validity of health DCEs has mainly focused on internal validity, such as whether model parameters have expected signs, and the ‘rationality’ of preferences defined by failure of non-satiation and/or lexicographic preferences (Bryan et al., 1998; Ryan, 1999; Ryan and Bate, 2001; Ryan et al., 2001; San Miguel et al.). It is important to note that rationality does not need non-satiation, and lexicographic preferences are not irrational (Lancsar and
Louviere, 2006). Work in health has also examined whether DCE results conform to axioms of completeness, monotonicity and transitivity (McIntosh and Ryan, 2002; Ryan and San Miguel, 2003; Lloyd, 2003). More research on axioms of transitivity and completeness, and tests of weak and strong axioms of revealed preference would be welcome. San Miguel et al. (2005) studied Sen’s expansion and contraction tests, and while not stated in their paper, such tests resemble a test of the weak axiom of revealed preference. Further work on non-compensatory preferences and/or other decision rules would also be welcome (Swait, 2001; Lloyd, 2003).

Unlike other areas in applied economics and marketing, there are few external validity tests of DCEs in health [Mark and Swait (2004) being an exception], possibly because it is harder to test external validity when a good or service of interest is publicly provided/financed. Yet, tests of external validity are possible, particularly for private goods; for example, medications are paid for out-of-pocket by patients in many health systems, providing opportunities to compare stated and revealed choices. Such research should receive priority because it would lend more credibility and confidence to DCE methods.

**Interpretation and policy application**

Most DCEs in health studied the relative importance of attributes, typically comparing size and significance of estimated taste parameters. However, choice model parameter estimates are confounded with the scale units of preference/utility measures of attribute levels and with error-variance differences within- and between-people, invalidating comparisons (Lancsar et al., 2007a). For relative attribute impacts to be compared, they must be measured on common comparable scales; five ways to do this were outlined by Lancsar et al. (2007a).

DCEs can be used to inform many policy and evaluation questions, and research has moved past estimating preferences and relative attribute importance to predict choice probabilities, useful for analysing likely uptake of health programmes (Hall et al., 2002; King et al., 2007), such as population health policies like cancer or immunisation screening programmes. Thus, DCEs may be useful in addressing the expanded remit of the UK NICE (National Institute for Health and Clinical Excellence) regarding the economic evaluation of population health initiatives. Similarly, use of a welfare-theoretic way to estimate compensating variation measures from DCEs (Lancsar and Savage, 2004a) is growing (e.g. Ryan et al., 2008; Ryan and Watson, in press), in lieu of calculating total WTP as a product of the sum of marginal rates of substitution and changes in attribute levels [the latter is theoretically inconsistent unless strict assumptions hold; see Lancsar and Savage (2004a, 2004b)].

Welfare-theoretic measures of value derived from DCEs can be used in CBA (Lancsar and Savage, 2004a; McIntosh, 2006). Indeed, a key advantage of using
DCEs to elicit monetary measures of benefit is that they can be used to value multiple options rather than valuing treatments/interventions one at a time. We are unaware of the use of DCE-derived monetary measures in CBA to date, but a framework for such use was suggested by McIntosh (2006), including use of a cost–benefit plane analogous to a cost-effectiveness plane. We also expect to see outcome measures from DCEs used in economic evaluation more generally, including measuring outcomes for use in cost-utility analysis (CUA); and utility weights derived from DCEs potentially can be used to calculate QALYs in CUA. Some work has been carried out on this topic (Hakim and Pathak, 1999; Viney et al., 2005; Ryan et al., 2006), but more is needed, such as comparisons with time tradeoff and standard gamble methods. There is also potential to elicit health and non-health related utility from DCEs for direct use in CUA, but more research is needed to determine feasibility (Lancsar and Louviere, 2008a).

DCEs can provide one component of priority-setting in general, and of priority-setting tools specifically, such as programme budgeting marginal analysis (Ruta et al., 2005). Similarly, DCEs have been included in RCTs (Longo et al., 2006; King et al., 2007), and consideration of their inclusion is likely to grow as DCEs become more common in economic evaluation. Decisions to include DCEs in RCTs should consider (at least) the specific research objectives of the DCE, whose preferences are relevant; implications for statistical power; justification of the costs of such an additional component; and the relation of DCE outcomes to other patient-reported outcomes collected in trials.

While initially DCEs may have been thought by some as a panacea, more than 15 years on from their first application in health, it is clear that, like all methods, they have potential limitations but also offer a number of advantages for use in economic (e)valuation and priority-setting in the health sector. Indeed, we see DCEs as complementary to other forms of preference elicitation and valuation in the health economist’s toolkit, and much of the work outlined in this section has highlighted the use of DCEs in conjunction with other SP and evaluation methods.

**Conclusion**

We provided a general history of DCEs and have reviewed their use in health economics. We briefly surveyed the current state-of-the-art, discussed key issues needing further research and emerging research trends. The literature on DCEs in health and other areas has grown rapidly over the past decade. While health applications have come a long way, we have argued that few were ‘best practice’, partly because ‘best practice’ has been something of a moving target. We have suggested ways to move various aspects of DCEs in the direction of best practice.
References


Fiebig, D. G., M. Keane, J. Louviere and N. Wasi (conditionally accepted), ‘The generalised multinomial logit model’, *Marketing Science*.


Louviere, J. and T. Engle (2006), ‘Confound it! That pesky little scale constant messes up our convenient assumptions!’, *Sawtooth Software Conference*, Sequim, USA.


Choice experiments in health: toward a brighter future


McIntosh, E. and J. Louviere (2002), Health Economics Study Group Meeting, Brunel University, January.


