PharmacoEconomics
DOI 10.1007/s40273-017-0506-4

PRACTICAL APPLICATION

# Discrete Choice Experiments: A Guide to Model Specification, Estimation and Software

**Emily Lancsar**[1] · **Denzil G. Fiebig**[2] · **Arne Risa Hole**[3]



**Abstract** We provide a user guide on the analysis of data (including best–worst and best–best data) generated from discrete-choice experiments (DCEs), comprising a theoretical review of the main choice models followed by practical advice on estimation and post-estimation. We also provide a review of standard software. In providing this guide, we endeavour to not only provide guidance on choice modelling but to do so in a way that provides a 'way in' for researchers to the practicalities of data analysis. We argue that choice of modelling approach depends on the research questions, study design and constraints in terms of quality/quantity of data and that decisions made in relation to analysis of choice data are often interdependent rather than sequential. Given the core theory and estimation of choice models is common across settings, we expect the theoretical and practical content of this paper to be useful to researchers not only within but also beyond health economics.

**Key Points for Decision Makers**

We provide a user guide on the analysis of data, including best–worst and best–best data, generated from discrete-choice experiments (DCEs), addressing the questions of 'what can be done in the analysis of DCE data' and 'how to do it'.

We provide a theoretical overview of the main choice models and review three standard statistical software packages: Stata, Nlogit and Biogeme.

Choice of modelling approach depends on the research questions, study design and constraints in terms of quality/quantity of data, and decisions made in relation to analysis of choice data are often interdependent rather than sequential.

A health-based DCE example for which we provide the data and estimation code is used throughout to demonstrate the data set-up, variable coding and various model estimation and post-estimation approaches.

## 1 Introduction

Despite researchers having access to ever-expanding sources and amounts of data, gaps remain in what existing data can provide to answer important questions in health economics. Discrete-choice experiments (DCEs) [1] are in demand because they provide opportunities to answer a range of research questions, some of which cannot otherwise be satisfactorily answered. In particular, they can provide insight into preferences (e.g. to inform clinical and

**Electronic supplementary material** The online version of this article (doi:10.1007/s40273-017-0506-4) contains supplementary material, which is available to authorized users.

✉ Emily Lancsar
Emily.Lancsar@monash.edu

[1] Centre for Health Economics, Monash Business School, Monash University, 75 Innovation Walk, Clayton, VIC 3800, Australia

[2] School of Economics, University of New South Wales, Sydney, NSW, Australia

[3] Department of Economics, University of Sheffield, Sheffield, UK

Published online: 03 April 2017

policy decisions and improve adherence with clinical/public health programmes or to understand the behaviour of key agents in the health sector, such as the health workforce, patients, policy makers, etc.), quantification of the trade-offs individuals are prepared to make between different aspects of healthcare (e.g. benefit–risk trade-offs), monetary and non-monetary valuation (e.g. valuing healthcare and/or health outcomes for use in both cost-benefit and cost-utility analysis and priority setting more generally) and demand forecasts (e.g. forecasting uptake of new treatments to assist in planning appropriate levels of provision).

DCEs are a stated-preference method that involve the generation and analysis of choice data. They are usually implemented in surveys; respondents are presented with several choice sets, each containing a number of alternatives between which respondents are asked to choose. Each alternative is described by its attributes and each attribute takes one of several levels that describe ranges over which the attributes vary.

Some reviews [2, 3] document the popularity and growth of such methods in health economics; others [4–6] provide guidance on how to use such methods in general. Unsurprisingly, given the detailed research investment needed to generate stated-preference data via DCEs, more detailed user guides on specific components of undertaking a DCE have also been developed, including the use of qualitative methods in DCEs [7], experimental design [8] and external validity [9]. A natural next large component of undertaking and interpreting DCEs to be addressed is guidance on the analysis of DCE data. This topic has recently received some attention: Hauber et al. [10] provide a useful review of a number of statistical models for use with DCE data. We go beyond that work in both scope and depth in this paper, covering not only model specification, which is the focus of the paper by Hauber et al. [10] (and within model specification we cover more ground), but also estimation, post-estimation and software.

We provide an overview of the key considerations that are common to data collected in DCEs, and the implications these have in determining the appropriate modelling approach, before presenting an overview of the various models applicable to data generated from standard first-best DCEs as well as for models applicable to data generated via best–worst and best–best DCEs. We discuss the fact that the parameter estimates from choice models are typically not of intrinsic interest (and why that is) and instead encourage researchers to undertake post-estimation analysis derived from the estimation results to both improve interpretation and produce measures relevant to policy and practice. Such additional analysis includes predicted uptake or demand, marginal rates of substitution, elasticities and welfare analysis. Coupled with this theoretical overview, we discuss how such models can be estimated and provide an overview of statistical software packages that can be used in such estimation. In doing so, we cover important practical considerations such as how to set up the data for analysis, coding and other estimation issues. We also provide information on cutting-edge approaches and references for further detail.

Many steps are involved in generating discrete-choice data prior to their analysis, including reviews of the relevant literature and qualitative work to generate the appropriate choice context and attributes and levels, survey design and, importantly, experimental design used to generate the alternatives between which respondents are asked to choose, and—of course—piloting and data collection. As noted, many of these 'front-end' steps have received attention in the literature and we do not discuss those topics here. Instead, we take as the starting point the question of how best to analyse the data generated from DCEs. Having said that, it is important to note that model specification and experimental design are intimately linked, not least because the types of models that can be estimated are determined by the experimental design. That means the analysis of DCE data is undertaken within the constraints of the identification and statistical properties embedded in the experimental design used to generate the choice data. For that reason, consideration of the types of models one is interested in estimating (and the content of this current paper) is important prior to creating the experimental design for a given DCE.

In providing this guide, we endeavour to not only provide guidance on choice modelling but to do so in a way that provides a 'way in' for researchers to the practicalities of data analysis. To this end, we refer throughout to and demonstrate the data set-up, variable coding, various model estimation and post-estimation approaches using a health-based DCE example by Ghijben et al. [11], for which we provide the data and Stata estimation code in the Electronic Supplementary Material (ESM). This resource adds an additional dimension that complements the guidance provided in this paper by providing a practical example to help elucidate the points made, and it can also be used as a general template for researchers when they come to estimate models from their own DCEs.

As such, the two main components of the paper are 'what can be done in the analysis of DCE data' and 'how to do it'. Given many (but not all) considerations in the analysis of DCE data are common across contexts in which choice data may be collected, we envisage the content of this paper being relevant to DCE researchers within and outside of health economics.

## 2 Choice Models

### 2.1 Introduction

Continual reference to the case study provided by Ghijben et al. [11] enables us to provide insights into some of the modelling decisions made in that paper as well as to make the associated data available as supplementary material to enable replication of all results produced in the current paper; some but not all of which appear in Ghijben et al. [11]. This carefully chosen example is broadly representative of the type of studies found in health economics and enables us to illustrate a relatively wide range of features of DCEs. Naturally, one example cannot provide an exhaustive coverage of issues likely to be faced by practitioners and, when appropriate, reference will be made to other applied work that supply templates for aspects that fall outside the scope of our case study.

Ghijben et al. [11] were motivated by the growing public health problem associated with atrial fibrillation and concerns of under-treatment. The study aimed to examine patient preferences for warfarin and new anticoagulants and motivated a decision problem where individuals faced with atrial fibrillation and an elevated risk of stroke needed to decide upon alternative treatments. These considerations motivated the development of the final choice task, an example of which is presented in Fig. 1. Sequences of such tasks were presented to respondents and form an integral part of the data collection. Again, we emphasize that getting to the stage of having a dataset amenable to analysis is a significant part of conducting a DCE and should not be underestimated. But this is not our focus, and interested readers should consult Johnson et al. [8], Dillman [12] and Tourangeau et al. [13] for details on the important topics of experimental and survey design.

It is useful to first introduce some terminology and to discuss a number of key features common to data collected in DCEs. In broad terms, the features are as follows:

A. *Discrete choices* On each choice occasion, respondents face a choice set containing two or more discrete and mutually exclusive alternatives. Respondents are then required to answer one or more questions reflecting their evaluation of these alternatives. In Fig. 1, respondents are required to first choose their most preferred alternative amongst the choice set of three options. They are then asked a follow-up question requiring them to choose the better of the two options remaining after their initial choice, which delivers a complete ranking of the three alternatives. Including just the first question is possibly the most common way to generate choice outcomes, and our discussion focuses on this case. However, the second type of question is an example that falls under the rubric of best–worst scaling (BWS) that is becoming increasingly popular because of the extra preference information provided at low marginal cost [14, 15].

B. *Choice sets* Choice sets contain two or more alternatives.[1] The choice set in our case study contains three alternatives, two referring to hypothetical drugs and one being a no-treatment option. Variants of such a structure include a status quo option so the investigator is determining which hypothetical alternatives would be attractive enough to make respondents switch from what they currently use; see Bartels et al. [16] and King et al. [17] for examples.[2] Where no-choice is a realistic alternative but is not provided as part of the choice set, the situation is referred to as a forced-choice problem. Including no-choice or status quo options usually adds realism to the choice task and is especially relevant in forecasting exercises and welfare analysis. The two hypothetical alternatives in Fig. 1 are fully described by their attributes, and the drugs are denoted by generic titles, 'drug A' and 'drug B'. They are said to be unlabelled alternatives. Sometimes it is more appropriate to provide a descriptive name for the hypothetical alternatives. For example, the choice set could include warfarin and a new oral anticoagulant such as dabigatran; the alternatives are now said to be labelled.

C. *Alternatives defined by attributes* Alternatives are defined by a set of attributes that are individually assessed by consumers in coming to an evaluation of the product as a whole [17]. The levels of the attributes are varied over choice occasions as part of the experimental design. Thus, the structure of these variables that figure prominently in subsequent analysis are under the control of the analyst. A good experimental design is one that ensures they deliver the best possible estimates and so problems prominent with revealed preference data, such as limited variation in key variables and multicollinearity, can be avoided in DCEs. It is true though that this comes with added responsibility on the part of analysts. For example, if there are interaction effects between attributes that are theoretically relevant, then it is necessary for the design to ensure that such effects are in fact identified.

D. *Repeated measures* The data have a panel structure with the same respondent providing multiple outcomes for a sequence of different choice occasions or

[1] This can include presenting a single profile and asking respondents to accept or reject it.

[2] In our case study, the status quo is no treatment; however, more generally, status quo and no treatment need not coincide.

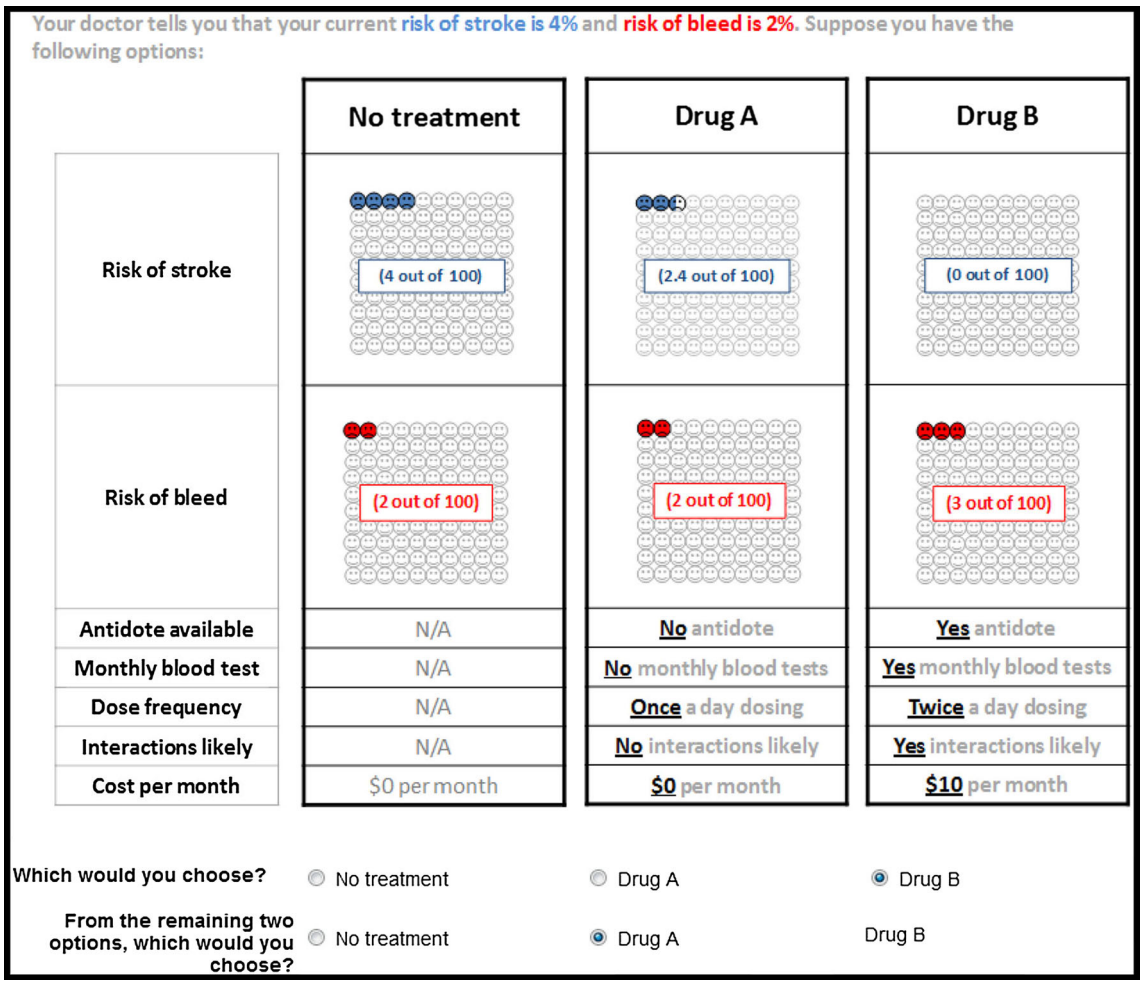


**Fig. 1** Example of a discrete-choice experiment choice set. Ghijben et al. [11]

scenarios. While asking respondents to answer more than one choice task is an economical way of gathering more information, it is clear that extra observations from the same respondent do not represent independent information. As in our case study, there are also examples where multiple outcomes are available for each scenario.

E. *Respondent characteristics* In a section of the survey instrument separate from the choice scenarios, personal characteristics of respondents are routinely collected. Different respondents may value different alternatives and attributes in different ways and so trying to capture these sources of preference heterogeneity with observable characteristics will typically form part of the analysis plan. In our case study, the medical history of respondents including any history of atrial fibrillation is potentially very relevant. While personal characteristics and relevant health history are natural inclusions, there is scope to collect and use less standard characteristics such as attitudinal variables [18–20].

F. *Context* Choices depend on the environment or context in which they are made [21]. In designing a DCE, the choice context plays a major role in making the hypothetical choice realistic. Context can also be manipulated as part of the experimental design by defining different contexts in which the choice is to be made and then allocating respondents to these context treatments and by including context variables as attributes. For example, our case study could be extended by allocating respondents to treatments that differed in the form or amount of information they were provided on atrial fibrillation or by including attributes of the cardiologist that respondents visited.

As well as providing an overview of typical DCE data and introducing some terminology, this initial discussion is important because several features of these data will eventually impact model specification.

### 2.2 Standard Discrete-Choice Models

It is natural to start with the classical multinomial logit (MNL) and its link to the random-utility model established by McFadden [22, 23]. This provides an opportunity to introduce most of the key specification and estimation issues and represents the baseline for most extensions to more sophisticated models and for research on the theoretical underpinnings of decision making in choice problems.

Assume the utility that respondent $i$ derives from choosing alternative $j$ in choice scenario $s$ is given by

$$U_{isj} = V_{isj} + \varepsilon_{isj}; \quad i = 1, \cdots, N; \quad s = 1, \cdots, S; \quad j = 1, \cdots, J; \tag{1}$$

where there are $N$ decision makers choosing amongst $J$ alternatives across $S$ scenarios. $V_{isj}$ represents the systematic or predictable component of the overall utility of choosing alternative $j$, and $\varepsilon_{isj}$ is the stochastic disturbance term representing characteristics unobservable by the analyst. We have data on the discrete choice $y_{is} = j$, which are then linked to the associated utilities by assuming the individual decision maker chooses alternative $j$ if it delivers the highest utility in comparison with the utility associated with all other alternatives in the choice set. Thus, we model the probability of choosing alternative $j$ as follows:

$$P_{isj} = \text{Prob}(y_{is} = j) = \text{Prob}(U_{isj} - U_{isl} > 0) \quad \forall l \neq j. \tag{2}$$

Equations 1 and 2 imply that the overall scale of utility is irrelevant in that multiplying both $V_{isj}$ and $\varepsilon_{isj}$ by a positive constant yields a different utility level but does not change the resultant choice. Consequently, the scale of utility needs to be normalized, which is equivalent to normalizing the variance of $\varepsilon_{isj}$.

Econometric analysis proceeds within this framework by making a number of assumptions and specification decisions. First, consider the distribution of the stochastic disturbance terms. Under the assumption that these are independently and identically distributed type-I extreme values, the probability of choosing $j$ takes the familiar MNL form:

$$P_{isj} = \frac{\exp(\lambda V_{isj})}{\sum_{l=1}^{J} \exp(\lambda V_{isl})} \tag{3}$$

where $\lambda$ is the scale parameter (inverse of the standard deviation of the disturbance). We will return to a further discussion of scale, but in a standard MNL model $\lambda$ cannot be separately identified and by convention is set to unity. Such normalizations should be familiar to anyone with knowledge of basic binary choice models such as logit and probit.

For modelling purposes, there is no compelling reason to prefer this specification for the disturbance distribution in preference to, say, normality that leads to a multinomial probit model. Historically, the preference for MNL arose because of the availability of a closed form solution for the probabilities as in Eq. 3, which leads to considerable computational advantages over multinomial probit where such representations of probabilities are not available. While computational considerations are today less of an issue, as we will see later, more complicated estimation problems can still benefit from having an MNL model as its base.

Turning to the specification of $V_{isj}$ as an initial starting point, one might consider a linear specification:

$$V_{isj} = \alpha_j + A_{isj}'\delta + Z_i'\gamma_j \tag{4}$$

where $A_{isj}$ is a vector of attributes describing alternative $j$, $Z_i$ is a vector of characteristics of the individual decision maker and $\alpha_j, \delta, \gamma_j$ are parameters to be estimated. Note the different sources of variation in the covariates. The attributes by design will typically vary over individuals, scenarios and alternatives while personal characteristics will only vary over individuals and will be constant over scenarios and alternatives. According to Eq. 2, only differences in utilities matter and so another generic specification issue is that characteristics of the individual have an impact on choice only to the extent that their associated parameters vary over alternatives, specified here as $\gamma_j$. Similarly, the alternative specific constants, $\alpha_j$, are specified to vary over alternatives. As there are $J - 1$ differences in utilities, it is also necessary to apply at least one normalization to the alternative specific constants and the parameters for the individual characteristics. This can be accommodated in a number of ways, but typically one alternative is set as the base and the associated parameters are normalized to zero. Because attributes do vary over alternatives, it is possible to estimate associated effects or preference weights that are not alternative specific. One could allow $\delta$ to vary over alternatives, but it is not necessary for the purposes of estimation.

Some discussions make a distinction between models of multinomial outcomes based on the structure of the regressors, calling a model with alternative-specific regressors a conditional logit model, one with case-specific regressors an MNL model and where there is a mixture of alternative-specific and case-specific regressors, as we have specified in Eq. 4, some authors call this a mixed model. They are essentially the same model, so such distinctions can lead to confusion. Thus, we simply call the model given by Eqs. 1–4 an MNL specification [24].

One way to interpret Eq. 4 is that the alternative specific constants vary by respondent characteristics. Faced with a

choice between the same alternatives, different individuals make different choices that can be predicted by differences in their observable characteristics. A natural extension to Eq. 4 also allows attribute weights to vary with respondent characteristics. Such heterogeneity can be captured by including interactions between attributes and individual characteristics. Decisions about these specification choices will depend on the subject matter of the particular problem and the research questions being considered.

As various models are introduced, we will supply associated estimation results produced using Stata and collect these in Table 1. Section 3 provides a comparison of estimates using alternative packages and more detail on estimation issues. The first column of results are for MNL, where we have used the first-best data and model B of Table 5 in Ghijben et al. [11] as the particular specification.

There are three alternatives, $j = N, A, B$, corresponding to no treatment and drugs $A$ and $B$. The no treatment choice parameters are normalized to zero; $\alpha_N = 0, \gamma_N = 0$. These constraints are necessary for identification. Ghijben et al. [11] impose two further constraints, neither of which is necessary for the purposes of identification, but both are sensible in this case. The first constraint, $\alpha_A = \alpha_B$, implies a single 'treatment' alternative-specific constant. Because the hypothetical alternatives are unlabelled, we would expect any preference for one drug over the other to be attributed solely to differences in attributes. In other words, if $A$ and $B$ were described by the same attribute levels, respondents would be expected to be indifferent between them. Conversely, in the case of labelled alternatives, it would not be prudent to impose such constraints. In such applications, alternative-specific constants would be capturing effects attributable to the label or brand over and above those captured by the attributes and hence these would be expected to differ across alternatives. Ghijben et al. [11] also imposed the constraint $\gamma_A = \gamma_B$. Once $\gamma_N = 0$ has been imposed, this second constraint is not necessary. It would be possible to specify alternative-specific attribute effects, for example, preferences over risk could differ depending on the treatment option, but specifying generic attribute effects is appropriate with these data.

In Table 1, the block effect, which is a dummy variable to control for the version of the survey to which the respondent answered, is written as an interaction with the treatment alternative-specific constant as are the personal characteristics in model $C$ of Table 3 in Ghijben et al. [11] (an exception is age, which is interacted with the risk attribute). This is operationally equivalent to allowing the associated parameters to vary between treatment and no treatment; depending on the options available in the software, this may be how the data need to be constructed for estimation to be undertaken (These and other estimation issues are addressed in Sect. 3).

While the specification has been shaped somewhat by the generic features of DCE data, a number of features have been overlooked in this basic MNL model. Simplicity of estimation and interpretation are among the main advantages of MNL, but these come at the cost of some

**Table 1** Estimates of standard discrete-choice models using data from Ghijben et al. [11]

| Variable | MNL | MXL | SH | G-MNL | ROL | MROL |
|---|---|---|---|---|---|---|
| Stroke risk | 0.698 (0.080) | 0.774 (0.088) | 0.886 (0.165) | 0.867 (0.117) | 0.503 (0.058) | 0.705 (0.076) |
| Bleed risk | 0.684 (0.080) | 0.742 (0.094) | 0.790 (0.135) | 0.816 (0.120) | 0.572 (0.066) | 0.767 (0.088) |
| Antidote | 0.066 (0.116) | 0.137 (0.133) | 0.108 (0.192) | 0.137 (0.147) | 0.149 (0.121) | 0.228 (0.146) |
| Blood test | −0.079 (0.075) | −0.088 (0.081) | −0.107 (0.099) | −0.120 (0.092) | −0.052 (0.059) | −0.038 (0.079) |
| Dose frequency | −0.064 (0.061) | −0.079 (0.065) | −0.132 (0.096) | −0.107 (0.078) | −0.014 (0.051) | −0.019 (0.067) |
| Drug/food interactions | −0.267 (0.083) | −0.314 (0.091) | −0.319 (0.097) | −0.338 (0.090) | −0.280 (0.078) | −0.385 (0.099) |
| Cost | −0.010 (0.001) | −0.011 (0.002) | −0.009 (0.002) | −0.011 (0.002) | −0.011 (0.001) | −0.014 (0.002) |
| Bleed risk[a] antidote | −0.337 (0.081) | −0.321 (0.094) | −0.320 (0.105) | −0.327 (0.100) | −0.213 (0.060) | −0.273 (0.077) |
| ASC[a] block | −0.589 (0.518) | −0.655 (1.533) | −0.810 (0.825) | −1.513 (0.575) | −0.216 (0.335) | −0.327 (0.656) |
| ASC (mean) | 0.926 (0.404) | 2.767 (0.966) | 1.824 (0.738) | 3.201 (0.538) | 1.043 (0.233) | 1.514 (0.431) |
| ASC (standard deviation) | | 3.052 (0.948) | | 3.178 (0.332) | | 2.438 (0.246) |
| $\tau$ | | | 0.961 (0.240) | −0.388 (0.109) | | |
| Log-likelihood | −944.3 | −786.1 | −884.7 | −781.8 | −1713.2 | −1377.9 |

Data are presented as estimate (standard error)

ROL/MROL estimates use the complete ranking data coming from the best–best choices of respondents, whereas all other columns of estimates only use the first best choices. All standard errors are cluster-robust, which allows for arbitrary correlation between the disturbance terms at the individual level. Estimation was undertaken in Stata 14.2

*ASC* alternative specific constant, *GMNL* generalised multinomial logit, *MNL* multinomial logit, *MROL* mixed rank ordered logit, *MXL* mixed logit, *ROL* rank ordered logit, *SH* scale heterogeneity model

restrictive assumptions that are clearly unrealistic in the context of DCEs.

The assumption of iid disturbances is especially problematic. We have already noted that the panel structure of the data is likely to induce correlation across choice occasions. A respondent in the case study with a preference for no treatment that is not captured by observable characteristics will carry that preference across choice occasions, inducing persistence in their choices. Such effects cannot be explicitly captured by any aspect of the current specification, but it is possible to adjust the standard errors to allow for clustering at the respondent level and this we have done in Table 1.

It has long been recognized that the independence of irrelevant alternatives (IIA) is a problematic aspect of the MNL. Proportional substitution across alternatives is a consequence of this model, irrespective of the actual data. Empirically, it may be a reasonable approximation in some settings, such as when all alternatives are generic. But in many DCE settings, especially those involving labelled alternatives, this is unlikely to be the situation. In the case study, it is highly likely that the impact of changing the cost of drug B is going to have a very different impact on the demand for drug A relative to the impact on the choice of the no treatment option. However, for MNL, we know beforehand that the predicted relative market shares would not change in response to a change in the cost of one drug. Thus, in situations where differential substitution patterns are likely, it is advisable to move to a more flexible specification.

It has already been suggested that attribute weights may vary with respondent characteristics. However, in modelling individual behaviour, unobserved heterogeneity is pervasive, implying a need to allow some or all of the parameters in Eqs. 3 and 4 to vary over individuals, even after controlling for variation explained by observable characteristics.

Rather than building up the model with extensions targeting separate issues, we will move to a very general model specification provided by the generalized MNL (G-MNL) developed in Fiebig et al. [25]. This is not the only model that could be chosen, and it is not without its critics [26], although these criticisms are more about interpretation than the model itself. It is a convenient choice in our discussion because it has the potential to capture all of the issues just raised. Moreover, it is a very flexible specification nesting several models often used in empirical applications and therefore provides a convenient framework for choosing between competing models. G-MNL is based on a utility specification that explicitly includes both individual-specific scale and preference heterogeneity. In other words, we allow $\lambda$, the scale parameter in Eq. 3, to vary by respondent: a form of heteroscedasticity. This allows for differential choice variability in that the errors are more important relative to the observed attributes for some respondents compared with others. Differences in scale are often interpreted as differences in choice consistency. In addition, the utility weights in Eq. 4 are assumed to be random coefficients that vary over respondents. G-MNL is written as follows:

$$U_{isj} = X'_{isj}\beta_i + \varepsilon_{isj} \tag{5}$$

$$\beta_i = \lambda_i\beta + \gamma\eta_i + (1-\gamma)\lambda_i\eta_i. \tag{6}$$

For notational convenience, all variables and parameters have been collapsed into single vectors $X$ and $\beta$.

If $\lambda_i = \lambda$, implying no scale heterogeneity (SH), G-MNL reduces to the mixed logit (MXL) specification, which has long been the most popular model used in DCE work; for examples, see Revelt and Train [27], Brownstone and Train [28], Hall et al. [29] and Hole [30]. After normalizing scale to unity, Eq. 6 becomes $\beta_i = \beta + \eta_i$ so that MXL is a random coefficient specification designed to capture preference heterogeneity, where $\eta_i$ represents random variation around the parameter means. There are error component versions of MXL [25] where the motivation comes from the need to induce correlation and heteroscedasticity across alternatives rather than from preference heterogeneity. Suppose only the alternative-specific constants are specified to be random. Assuming they are correlated provides a convenient way to avoid IIA and allow more flexible substitution between choices. It also provides a means to capture dependence due to the panel structure because $\eta_i$ varies over individuals but it is assumed to be fixed over choice occasions. This then induces a positive correlation across choice occasions and represents a typical baseline specification common in panel analyses. Importantly, it will provide estimated standard errors that better reflect the nature of the data. Such a specification has been estimated, and the results are denoted by MXL in Table 1.

If $\eta_i = 0$ so that there is no preference heterogeneity, G-MNL reduces to a model where $\beta_i = \lambda_i\beta$. Allowing only for SH indicates how this specification is observationally equivalent to a particular type of preference heterogeneity in the utility weights, an observation that led Louviere et al. [31] to be critical of the standard MXL model. The estimation results for this model are denoted by SH in Table 1. While conceptually there is a difference between SH and preference heterogeneity, they are intrinsically linked and, hence, in practice it is difficult to disentangle the two empirically [26]. In particular, rather than making this distinction, one could simply motivate the G-MNL as a flexible parametric specification for the distribution of heterogeneity.

Restricting $\gamma$ to zero implies $\beta_i = \lambda_i(\beta + \eta_i)$, whereas $\gamma$ equal to one implies $\beta_i = \lambda_i\beta + \eta_i$. In these two variants of

G-MNL, it is either the random coefficients or just their means that are scaled. Both of these are sensible alternatives, although the choice between them may not be a priori obvious. Freely estimating $\gamma$ allows more flexibility in how the variance of residual taste heterogeneity varies with scale. The estimation results denoted by G-MNL in Table 1 are for a specification that simply combines the features of the current MXL and SH specifications with $\gamma = 0$. The full G-MNL model provides flexibility in how SH and taste heterogeneity are combined but can involve a large number of parameters, especially when there is a large number of alternatives and their random coefficients are assumed to be correlated. Given the relatively small sample size in the case study, this parsimonious specification is a sensible choice here.

The specification is completed by choosing distributions that capture the individual heterogeneity. Preference heterogeneity is often specified to follow a multivariate normal distribution. This is what has been assumed in the results used to generate MXL and G-MNL in Table 1. In principle, the choice of distribution is flexible, but—in practice—normality is by far the most common choice. Lognormality is another popular option used when restricting coefficient signs. SH in the SH and G-MNL models is assumed to be as follows:

$$\lambda_i = \exp\left(\bar{\lambda} + \tau v_i\right) \tag{7}$$

where $v_i \sim N(0, 1)$ and $\bar{\lambda}$ is a normalizing constant required to ensure identification of $\lambda_i$. Note, the GMNL results in Table 1 include scaling the alternative specific constant (ASC); in other situations, it may be problematic to do so; see Fiebig et al. [25] for further discussion. The additional parameter $\tau$ provides a measure of SH. If $\tau = 0$, the G-MNL model reduces to a standard MXL specification. The work of Fiebig et al. [25] highlights the empirical importance of accommodating this extra dimension of heterogeneity. One attraction of this specification of SH is a considerable amount of flexibility with the addition of only one parameter. In principle, researchers have considerable flexibility in choosing these distributions, but in practice most are confined to what options are offered in available software (Again these are issues are addressed in Sect. 3).

We stress that there are different approaches to model specification. An alternative to specifying random coefficient models that is quite popular in applications is to assume that there are a finite number of types where parameters vary within but not across types. So, modelling heterogeneity is again the motivation but finite mixture or latent class models result; see Hole [30] for an example. Keane and Wasi [32] provide a comparison of the two approaches in terms of fit and, while they do not proclaim a clear winner, they do suggest that G-MNL performs well in comparison with finite mixture models in part because the former tends to be more parsimonious.

### 2.3 Best–Worst and Best–Best Discrete-Choice Models

The term BWS has been used somewhat loosely in the literature. It is in fact a generic term that covers three specific cases: (1) best–worst object scaling, (2) best–worst attribute scaling and (3) best–worst DCEs (BWDCE). All three involve asking respondents to choose the best and worst (most and least preferred) from a set of three or more items. In reverse order, respondents choose best and worst between alternatives in (3); between attribute levels within a single alternative or profile in (2) and between whole objects (or sometimes statements, principles, etc.) that are not decomposed into attributes in (1). See Lancsar et al. [14] and Louviere et al. [15] for further description and comparison of the three cases. Here, we focus on BWDCEs for two reasons. First, they are the form of BWS closest to traditional DCEs, or can be thought of as a specific type of DCE. Second, the models we discuss here for BWDCE data can readily be applied to the other two types of BWS.

Like standard DCEs, respondents make repeated choices between alternatives offered in choice sets, each described by a number of attributes. However, BWDCEs are designed to elicit extra preference information per choice set by asking respondents not only to choose the best option but also to sequentially choose the worst option, potentially followed by choice of best of the remaining options and so on until an implied preference ordering is obtained over all alternatives in a set. For a choice set containing $J$ alternatives, respondents can be asked to make (and the analyst build models based on) up to $J - 1$ sequential best and worst choices. At a minimum, a BWDCE doubles the number of observations for analysis (or more if more alternatives are included per set and additional best and worst questions are answered), which in turn can be used to increase the statistical efficiency of the choice models or reduce sample sizes for a given target number of observations [14]. By providing a higher number of degrees of freedom for analysis, the extra preference data obtained via BWDCEs also opens up new research avenues such as the estimation of individual-level models [35], something generally not possible with the amount of data collected per person in a standard DCE.

More recently, Ghijben et al. [11] introduced a variation of BWDCE, namely a best–best DCE in which best is chosen from the full choice set followed by repeated choice of best (instead of worst) from the remaining alternatives until a preference order is obtained over all alternatives, and used this elicitation process in their study. Like a BWDCE, this increases the number of observations collected per choice set but does so without

asking respondents to swap to a new mental task of choice of worst.

We discuss three ways to analyse best–worst data that account for varying amounts and composition of preference information. The first is to simply harness the first (best) choice in each choice set, ignoring the additional choice data using models outlined in Sect. 2.2 (as done by Lancsar et al. [33] and Fiebig et al. [34]). Alternatively, the additional preference information obtained from a BWDCE can be used to estimate discrete-choice models by noting that the best and worst choice questions produce an implied rank order over alternatives, which can be modelled with rank ordered logit (ROL) (e.g. Lancsar and Louviere [35] and Scarpa et al. [36]). ROL [37–39] models the probability of a particular ranking of alternatives as the product of MNL models for choice of best. For example, the ranking of three alternatives $A > B > C$ is modelled as the product of the (MNL) probability of choosing $A$ as best from the set ($A$ $B$ $C$) times the probability of choosing $B$ as best from the remaining alternatives ($B$ $C$).

$$\begin{aligned}\Pr(\text{ranking } A,B,C) &= \Pr(A \text{ is1st best}) * \Pr(B \text{ is 2nd best}) \\ &= \frac{\exp(V_A)}{\sum_{j=A,B,C}\exp(V_j)} * \frac{\exp(V_B)}{\sum_{j=B,C}\exp(V_j)}\end{aligned} \tag{8}$$

Subscripts are omitted for notational brevity. In using ROL to estimate the implied preference order from a BWDCE, the best–worst structure used to generate that order is ignored since ROL assumes best (not worst) is chosen from successively smaller choice sets. In contrast, the ROL matches exactly the data-generation process of a best–best DCE in which best is chosen from successively smaller choice sets. The data are modelled using ROL in Table 1.

The sequential best–worst MNL (SBWMNL) model developed in Lancsar and Louviere [14, 35] directly models the series of sequential best and worst choices made in each choice set as the product of MNL models. Using the aforementioned example of a choice set containing three alternatives ($A$ $B$ $C$), the probability of observing the preference order $A > B > C$ is modelled as the (MNL) probability of choosing $A$ as best from the set ($A$ $B$ $C$) times probability of choosing C as worst from the remaining alternatives ($B$ $C$), which can be expressed as follows:

$$\begin{aligned}&\Pr(\text{best worst ordering A, B, C}) \\ &\quad = \Pr(A \text{ is best}) * \Pr(C \text{ is worst}) \\ &\quad = \frac{\exp(V_A)}{\sum_{j=A,B,C}\exp(V_j)} \times \frac{\exp(-V_C)}{\sum_{j=B,C}\exp(-V_j)}\end{aligned} \tag{9}$$

Here, the best–worst ordering of the three alternatives is represented as the two choices made by respondents per choice set in the BWDCE and the deterministic part of utility of choosing an alternative as worst is modelled as the negative of the deterministic utility of choosing that alternative as best. SBWMNL models the choice data in the way they were generated so that the worst choice is modelled in the second MNL model in Eq. 8 and the composition of the denominator reflects the actual choice sets considered by respondents associated with each choice set in the sequence.

Equations 7 and 8 can both be generalized to account for the types of heterogeneity discussed in Sect. 2.2. For example, Lancsar et al. [14] demonstrated how mixed logit, heteroscedastic logit and G-MNL versions of ROL and SBWMNL could be estimated. The mechanics of doing so is made very straightforward because (as we discuss in Sect. 3) both ROL and the SBWMNL models can be estimated by 'exploding' the data into the implied choice sets and then estimating MNL on the exploded data. As such, it is straightforward to estimate any of the models discussed earlier on rank or best–worst data. In the final column of Table 1, we present the estimates for the mixed ROL (MROL) that reproduces the specification B results from Table 3 in Ghijben et al. [11]. Thus, these results are directly comparable to the MXL estimates in Table 1 that just uses the first best choice.

### 2.4 Post-Estimation

Good econometric practice suggests one should conduct various robustness checks to ensure the internal validity of the estimated model. When dealing with revealed preference data, a constant threat is omitted variable biases, meaning that effects of interest may be very sensitive to the variables not included in the model. This is much less of a concern with stated-preference data where the effects of interest are typically associated with the attributes that are created as part of the experimental design, meaning there is typically no correlation between attributes and no reason to believe they will be correlated with respondent characteristics. For omitted variable biases to arise in stated preferences, attributes would need to be omitted from the design that lead respondents to change their evaluation of included attributes because of expectations about the relationship between omitted and included attributes. This would lead to biased estimates of coefficients of included attributes, and these biases could depend on respondent characteristics. But such problems can readily be minimized at the design stage and is an argument for avoiding simple designs with minimal numbers of attributes.

What is a potential threat are features of the design such as when respondents are assigned to different versions of the survey. In the case study, the authors included a dummy variable to control for block effects. From Table 1, we see that these effects are never statistically significant at any conventional level and so, at least in this dimension, one should be confident about the results.

Often issues of model choice can be resolved by testing restrictions associated with nested versions of more general models. Because maximum likelihood is the basis for estimation, likelihood ratio tests can easily be conducted for this purpose. For example, in Table 1, MNL, MXL and SH are nested within G-MNL. It is wise to remember that, except for MNL, the log-likelihood function is simulated rather than known exactly and so is subject to simulation noise. In the case of non-nested models, such as a comparison between MXL and SH or between MROL and MXL, information criteria such as Akaike information criterion (AIC) and Bayesian information criterion (BIC) may be used to discriminate between alternative models.

While model fit may provide some guidance in choice of models, often what is more important is the specific research question being considered and the specific features of the data being used and how they map into the model being considered. For example, if data are being pooled across very different subsamples of respondents, SH would be an obvious concern; see, for example, Hall et al. [29]. If one is simulating the impact of the introduction of a new product or treatment as in Fiebig et al. [34], then it would be prudent to avoid MNL and allow for flexible substitution patterns.

Section 3.4 discusses other post-estimation issues, such as how the estimated model can be used to generate measures of marginal willingness to pay (mWTP) and carry out predictive analyses.

## 3 Software and Estimation

### 3.1 Discrete Choice

This section provides an overview of software for estimating the models described in the previous section, summarised in Table 2. The focus is on general statistics/econometrics packages with built-in commands for estimating discrete-choice models,[3] rather than programming languages that require user-written code.[4]

#### 3.1.1 Nlogit

Nlogit (www.limdep.com/products/nlogit) is an extension of the Limdep statistical package. It has a very comprehensive set of built-in commands for estimating discrete-choice models, and can be used to estimate all of the models covered in Sect. 2. It has various post-estimation routines for generating predicted probabilities, performing simulations and calculating elasticities. Nlogit is relatively easy to use and comes with a comprehensive manual as a PDF.

#### 3.1.2 Stata

Stata (www.stata.com) is a general statistics package that offers a broad range of tools for data analysis and data management. While it has fewer built-in commands for estimating discrete-choice models than Nlogit, there is a range of user-written commands freely available that can be used to implement the methods covered in Sect. 2 [40–43]. It has routines for generating predicted probabilities, and simulations can be performed and elasticities calculated by using the generated probabilities. Like Nlogit, Stata is relatively easy to use and comes with a comprehensive manual as a PDF. User-written commands are often documented in articles published in *The Stata Journal* (www.stata-journal.com).

#### 3.1.3 Biogeme

Unlike Stata and Nlogit, which are general statistical packages, Biogeme (biogeme.epfl.ch) is specifically created for estimating discrete-choice models.[5] It also stands out for being the only package of the three that is free; both Stata and Nlogit require the user to pay a licence fee. Biogeme is capable of estimating MNL models with both linear and non-linear utility functions and with random coefficients, which means that all of the models covered in Sect. 2 can be implemented. It also has a routine for performing simulations (biosim). While, in our experience, Biogeme is somewhat less easy for beginners to use than Stata and Nlogit, the documentation is comprehensive and has helpful examples. Since Biogeme requires a somewhat higher initial time investment, it is recommended in particular for more advanced users who wish to go beyond standard model specifications. As Biogeme has fewer built-in commands for data management than Stata and Nlogit, it will often be necessary to use an alternative software package to set up the data in the form required by Biogeme.

[3] Our overview is not exhaustive, as other software packages capable of estimating some of the discrete-choice models in our review are available. However, the three packages we have reviewed are among the most commonly used for estimating these models.

[4] This implies we will not cover software such as Gauss, Matlab and R, despite there being excellent routines written in these packages for estimating, for example, mixed logit models. A prominent example is Kenneth Train's codes for mixed logit estimation (http://eml.berkeley.edu/~train/software.html), which served as inspiration for many of the routines later introduced in other statistical packages.

[5] Two versions of Biogeme are available: BisonBiogeme and PythonBiogeme. We focus on BisonBiogeme, which is designed to estimate a range of commonly used discrete-choice models.

**Table 2** A summary of the modelling capabilities of the main software packages covered in the review

| Software package | MNL | MXL | Latent class | MXL Bayesian | G-MNL |
|---|---|---|---|---|---|
| Nlogit | ✔ | ✔ | ✔ | | ✔ |
| Stata | ✔ | UW | UW | UW | UW |
| Biogeme | ✔ | ✔ | ✔ | | ✔ |

This is not an exhaustive list; all of the packages have options for estimating other discrete-choice models not covered in this review. The packages differ in terms of which distributions are supported for the random coefficients in the mixed logit routines, with Nlogit having the widest selection of distributions. GMNL can be fit in Biogeme by exploiting the option for specifying non-linear utility functions; however, it is less straightforward to do than in the other two packages. Scaled ASCs are the default GMNL option in Stata and Nlogit and can be done in Biogeme. However, this can be potentially problematic depending on the context and model and we suggest testing with and without scaling the ASC (in our case, it made little difference)

*ASC* alternative specific constant, *GMNL* generalised multinomial logit, *MNL* multinomial logit, *MXL* mixed logit, *UW* user-written

To sum up, Nlogit, Stata and Biogeme are all good options for estimating discrete-choice models. We would argue that there is no 'one size fits all': no package strictly dominates the others, and it will therefore be up to the individual user to choose the package that best suits their needs. As mentioned, Biogeme has the advantage of being free and very powerful, but it requires a somewhat greater time investment on the part of the user to learn how to use it effectively. Nlogit has a very comprehensive set of built-in commands that cover most, if not all, of the models that the majority of DCE analysts would want to estimate. Stata has a less comprehensive set of built-in commands, but has user-written routines that cover the most commonly used models. Both Nlogit and Stata have the advantage that all data processing and cleaning can be done in the same package that is used to run the analysis. All three packages have active online user group discussion forums where queries are typically answered quickly. The forum archives are searchable and contain a wealth of useful information in the form of past questions and answers, so it is typically worth spending some time searching the archives before posting a new question.

## 3.2 Estimation of Discrete-Choice Models

### *3.2.1 Data Setup*

Before proceeding to the estimation stage, the analyst needs to organise the data in the way required by the estimation software. In general, the data can be organized in two ways: 'long form' (see Supplementary Appendix 1) and 'wide form' (see Supplementary Appendix 2):

- Long form, which is the data structure required by Stata and Nlogit,[6] implies that the dataset has one row per alternative for each choice scenario that the decision makers face. Thus, with $N$ decision makers choosing amongst $J$ alternatives across $S$ scenarios, the dataset will have $N \times J \times S$ rows. The dependent variable is coded 1 for the chosen alternative in each scenario and 0 for the non-chosen alternatives.
- Wide form, which is the data structure required by Biogeme, implies that the dataset has one row for each choice scenario that the decision makers face. The dataset will therefore have $N \times S$ rows. In this case, the dependent variable is coded $1, \ldots, J$, indicating the chosen alternative. Each design attribute will have $J$ associated variables, containing the level of the attribute for the respective alternative. This contrasts with the long form structure, where there is only one variable per design attribute.

Both Stata and Nlogit have built-in commands for transforming the dataset from long to wide form, and vice versa. For convenience, the example dataset is available as ESM in both long and wide form.

When the data are in long form, ASCs can be defined as a dummy variable that is equal to one in the row corresponding to the relevant alternative and zero otherwise. Alternative-specific coefficients can then be estimated by interacting the ASCs with the desired attribute(s) and including the interactions in the model. In Biogeme, such effects are specified by explicitly defining the utility function of the different alternatives, an option that is also available in Nlogit. More generally, categorical attributes and covariates can be coded as dummy variables or effects coded, either being appropriate as long as interpreted appropriately [44, 45]. Indeed, even for continuous variables (e.g. price), it can often prove useful to initially treat the levels as categorical in exploratory testing in order to plot the coefficients to help inform choice of functional form for the continuous variable.

[6] Nlogit also optionally allows the data to be organized in wide form, although the manual suggests that long form is typically more convenient.

#### 3.2.2 Other Estimation Issues

Implementing the models presented in Sect. 2 in practice requires the analyst to make a number of choices at the estimation stage.[7] As we show, these choices can impact on the results, and it is therefore recommended to carry out a sensitivity analysis to examine the robustness of the findings.

*Starting values* Estimating the parameters in the model involves maximising a non-linear log-likelihood function, and the maximisation process requires the user to provide an initial guess of the parameter values.[8] The software package then searches for improvements in the log-likelihood iteratively by changing the values of the parameters using an optimisation algorithm. In the case of the MNL model, the choice of starting values typically does not matter in practice, as the MNL log-likelihood function has a single maximum. The algorithm will therefore find the maximum even if the starting values are far from the values that maximize the log-likelihood. However, in the case of models such as MXL and G-MNL, matters are less simple. For those models, the log-likelihood may have several optima, of which only one is the overall (global) optimum that we seek to identify. Starting from a set of parameter values far away from the global optimum may lead the algorithm to identify one of the inferior local optima, at which point the algorithm will declare convergence as it cannot distinguish between local and global optima. Only the parameter values associated with the global optimum have the desirable properties of maximum likelihood estimates, and it is therefore recommended to investigate the sensitivity of the results to a different choice of starting values. Hole and Yoo [46] discuss these issues in the context of the G-MNL model. Czajkowski and Budziński [47] find that increasing the number of simulation draws improves the chance of the algorithm converging to the global optimum.

*Simulation draws* Another issue that the analyst needs to be aware of when estimating MXL and G-MNL models is that the log-likelihood function must be approximated using simulation methods, as it cannot be calculated analytically. Simulation methods involve taking a large number of random draws, which represent the distribution of the coefficients at the current parameter values. As the draws are generated by a computer, they are not truly random but instead created using an algorithm designed for the purpose of generating draws, which have similar properties to random draws. The analyst needs to decide how many draws to use to approximate the log-likelihood function and which method to use to generate the draws. Regarding the number of draws, there is a trade-off between accuracy and estimation time; a large number of draws gives a better approximation of the true log-likelihood function but slows down the estimation process. It is therefore common to start with a relatively small number of draws at the exploratory stage, for example using the default setting in the software.[9] It is then strongly advisable to check for the stability of the final solution to be reported by increasing the draws. The number of draws required to stabilize the results will depend on the model specification; typically, a larger number of draws is needed if there are more random coefficients in the model. The number of draws required is also related to the method chosen to generate the draws. For example, in the context of mixed logit estimation, 100 Halton draws have been found to be more accurate than 1000 pseudo-random draws [48, 49]. For this reason, Halton draws are often used when estimating MXL and G-MNL models.[10]

Table 3 presents the results from estimating a simplified version of the MXL model described in Sect. 2 (model A from Ghijben et al. [11]) in Nlogit, Stata and Biogeme. The starting values are set at the default values in Nlogit and Stata, and the Biogeme starting values are set to be identical to the Stata default values.[11] While it can be seen that there are no qualitative differences between the results—the coefficients have the same sign and significance and the point estimates are similar—they are not exactly identical. This is in spite of using the same number of simulation draws (500) and the same method for generating the simulation draws (Halton).[12] However, as long as the results do not differ to the extent that it has an impact on the substantive implications of the findings, this should not give much cause for concern.

### 3.3 Estimation of Best–Worst and Best–Best Models

Estimation of choice models harnessing just the first best choice from best–worst or best–best data proceeds as

[7] Interested readers are referred to chapters 8–10 in Train [49] for more information about the issues covered in this section.

[8] Both Nlogit and Stata will use a default set of starting values unless explicitly specified by the user, whereas Biogeme requires the user to specify the starting values.

[9] The default number of draws is 100 in Nlogit, 50 in Stata and 150 in Biogeme.

[10] In models with several random coefficients, alternative approaches such as shuffled or scrambled Halton draws [50] or Sobol draws [51, 52] are sometimes used to minimize the correlation between the draws, which can be substantial for standard Halton draws in higher dimensions. See chapter 9 in Train [49] for a discussion.

[11] Nlogit and Stata's default starting values are the MNL parameters for the means of the random coefficients and 0 (Nlogit)/0.1 (Stata) for the standard deviations.

[12] Differences can still arise, for example because the optimization algorithms differ in the three packages, subtle differences in terms of how the Halton draws are generated and different starting values (in this case Stata/Biogeme vs Nlogit).

**Table 3** Results from estimating a MXL model in the different software packages

| Variable | Nlogit | Stata | Biogeme |
|---|---|---|---|
| Stroke risk | 0.706 (0.060) | 0.706 (0.060) | 0.706 (0.060) |
| Bleed risk | 0.578 (0.049) | 0.578 (0.049) | 0.578 (0.049) |
| Antidote | 0.600 (0.081) | 0.600 (0.081) | 0.600 (0.081) |
| Blood test | −0.082 (0.077) | −0.082 (0.077) | −0.082 (0.077) |
| Dose frequency | −0.089 (0.077) | −0.089 (0.077) | −0.089 (0.077) |
| Drug/food interactions | −0.340 (0.079) | −0.340 (0.079) | −0.340 (0.079) |
| Cost | −0.012 (0.001) | −0.012 (0.001) | −0.012 (0.001) |
| ASC x block | −1.051 (0.862) | −1.003 (0.856) | −1.000 (0.854) |
| ASC (mean) | 3.108 (0.759) | 3.086 (0.750) | 3.090 (0.742) |
| ASC (SD) | 3.102 (0.492) | 3.069 (0.477) | 3.050 (0.478) |
| Log-likelihood | −791.03 | −790.91 | −790.85 |

Data are presented as coefficient (standard error) unless otherwise indicated

The following versions were used for estimation: Stata 14.2, NLOGIT 5 and Biogeme 2.0

*ASC* alternative specific logit, *SD* standard deviation

outlined in Sect. 3.1. Estimation of ROL in Stata involves the rologit command. In Nlogit, ROL models can be estimated using the rank as the dependent variable and adding 'ranks' to the usual model syntax. In Biogeme, the data need to be exploded manually (see Supplementary Appendix 3 in the ESM). Indeed, an attractive property of both ROL and SBWMNL is that they can be estimated using standard MNL (or extensions such as MIXL, G-MNL, etc.) after the data have been set up appropriately. In fact, ROL is also known as 'exploded logit' because, drawing on the IIA property of MNL models, it can be estimated by exploding the data from each choice set into statistically independent choice subsets. For a choice set with $J$ alternatives, the data can be expanded into $J - 1$ sub choice sets. For example, for a ranking over $J = 3$ alternatives, the data can be exploded into two sub choice sets. The first contains three rows of data representing the three alternatives contained in the original choice set with the dependent variable equal to 1 for the alternative ranked first (chosen as best) and 0 for the remaining alternatives, which is identical to data set-up for standard first best choice model. The second sub choice set identifies best from the remaining two alternatives and contains two rows of data pertaining to the two alternatives not ranked first with the dependent variable equal to 1 for the alternative ranked second (chosen as best from the two on offer) and 0 for the remaining alternative. So, for each original choice set containing three alternatives, there are five rows of data.[13] Once the rankings are exploded in the dataset to the implied choices made in each of the subsets, the ROL parameters can be estimated using a traditional MNL model (or extensions of the MNL model) from the expanded choice data. Indeed, prior to ROL routines being programmed in software packages, this was the standard way to estimate a ROL. A good check that the data have been exploded correctly before moving on to more sophisticated models accounting for unobserved heterogeneity, etc. is to run an MNL (e.g. via the clogit or asclogit command in Stata) on the exploded data and then run an ROL on the un-exploded data (e.g. using rologit in Stata). The results should be identical in all decimal places.

In all three software packages (Stata, Nlogit and Biogeme), the data need to be exploded to estimate SBWMNL models. Like the ROL model, estimation of the SBWMNL model draws on the IIA property and exploits the additional preference information obtained in each choice set in a BWDCE, expanding the data in a similar but slightly different way. Again, for a choice set with $J$ alternatives, the data can be exploded into $J - 1$ sub choice sets. So data from a choice set containing three alternatives from which best and worst are chosen are expanded into two sub choice sets. The first contains three rows of data corresponding to the three alternatives presented in the original choice set with the dependent variable equal to 1 for the alternative chosen as best, and 0 for the remaining alternatives. The second sub choice set contains two rows of data representing the two alternatives not chosen as best in the full choice set with the dependent variable equal to 1 for the alternative chosen as worst and 0 for the remaining alternative. Thus, again for each original choice set containing three alternatives, there are five rows of data. In addition, for the sub choice set in which worst is chosen, the utility of worst is scaled to be the negative of the utility of best, which in practice means multiplying the data for the alternatives in the second sub choice set by −1. Parameters can then be estimated using MNL (or its extensions) on the exploded choice data. Code for all

[13] Applying this procedure modifies the data from the standard set-up in Supplementary Appendix 1 to the exploded set-up in Supplementary Appendix 3.

estimation contained in this paper is provided in Supplementary Appendix 4 (ESM).

### 3.4 Post-Estimation

As discussed in Sect. 2.4, issues of model choice can be resolved by testing restrictions associated with nested versions of more general models. Such tests can easily be carried out using the tools available in Nlogit and Stata for performing Wald tests. In all packages, an alternative approach is to carry out a likelihood ratio test using the reported simulated log likelihood values for the restricted and unrestricted models.

Information criteria such as AIC and BIC may be used to discriminate between alternative models that are not nested. Apart from the examples mentioned in Sect. 2.4, this can be useful when assessing the goodness of fit of MXL models with different distributions for the random coefficients, for example. The information criteria can be readily calculated in all packages using the reported simulated log likelihood values, along with the relevant information regarding the number of parameters in the competing models and (in the case of BIC) the sample size.

It is worth re-iterating that while model fit may provide some guidance in choice of models, often what is more important is the specific research question being considered and the specific features of the data being used. Rather than choosing one preferred specification, it is typically better to report the output of interest (such as mWTP measures, predictive analyses) for a range of model specifications and compare and contrast the results, as demonstrated below.

#### *3.4.1 Marginal Willingness to Pay Measures*

Calculating mWTP measures is a convenient and useful way to compare attribute estimates. mWTP can be derived as the marginal rate of substitution between attribute $X^k$ and cost ($C$):

$$\text{mWTP}_{X^k} = -\frac{\text{MU}_{X^k}}{\text{MU}_C} \tag{10}$$

where $\text{MU}_{X^k}$ and $\text{MU}_C$ are the marginal utilities of attribute $X^k$ and cost, respectively. When the utility function is specified to be linear in parameters, the marginal utility of an attribute is equal to its coefficient, which means that mWTP is given by the negative of the ratio of the coefficients for attribute $X^k$ and cost. In Table 4, we use the MNL, MXL and MROL results from Table 1 to produce mWTP estimates for the most important attributes, stroke risk and bleed risk. The latter replicate the results in Table 4 in Ghijben et al. [11].

The MNL results indicate that, for example, the WTP for a 1%-point reduction in the stroke risk is about 69 Australian dollars (AUD) per month. The remaining estimates can be interpreted in an analogous way. The estimates are reasonably stable across models, with the MNL and MXL estimates being especially close, whereas the MROL stroke risk mWTP is somewhat of an exception. If we refer back to the actual parameter estimates used in these calculations, those for MNL are systematically smaller in magnitude, a scaling effect, but those for MXL and MROL are very similar, as we would expect.

While mWTP measures are straightforward to calculate when the utility function is linear in parameters, routines for obtaining confidence intervals using either the delta method or parametric (Krinsky–Robb) or non-parametric bootstrapping [53] are useful since a measure of the precision of the estimates should always be reported. Such routines are available in both Nlogit and Stata. In some cases, the assumption of linearity is inappropriate, as a researcher may want to allow the marginal utility of a change in an attribute to depend on the level of the attributes (e.g. due to interactions or non-linear functional form; for an example, see Lancsar et al. [33]). In such cases, the calculation of mWTP is slightly more involved, but we can still use general routines for calculating non-linear combinations of parameters in Stata and Nlogit to obtain point estimates and measures of precision.

Calculating mWTP measures following the estimation of a model with random coefficients, such as MXL or G-MNL, can be more complicated depending on the model specification. If both the attribute coefficient and the cost coefficient are fixed, as in our examples, the calculation is the same as for the MNL model. If the attribute coefficient is normally distributed and the cost coefficient is fixed, which is a common specification, the mean mWTP is

**Table 4** Selected estimates of marginal willingness to pay for a subset of attributes using data from Ghijben et al. [11]

| Attribute | MNL | MXL | MROL |
|---|---|---|---|
| Stroke risk | 68.37 (44.24–92.51) | 70.53 (44.31–96.75) | 50.48 (33.68–67.28) |
| Bleed risk (without antidote) | 67.02 (42.29–91.75) | 67.66 (41.09–94.23) | 54.87 (35.70–74.05) |
| Bleed risk (with antidote) | 34.02 (18.55–49.50) | 38.36 (21.93–54.79) | 35.70 (21.70–48.99) |

Data are presented as estimate (95% confidence interval)

Marginal willingness to pay per month in Australian dollars for a 1 percentage point reduction in absolute risk

*MNL* multinomial logit, *MROL* mixed rank ordered logit, *MXL* mixed logit

**Table 5** Comparison of predictions for each alternative in response to changes in selected attributes using data from Ghijben et al. [11]

| Model and scenario | No treatment | Drug A | Drug B |
|---|---|---|---|
| MNL | | | |
| Baseline | 0.165 (0.079–0.317) | 0.417 (0.341–0.461) | 0.417 (0.341–0.461) |
| Increase cost of $50 for drug B | 0.198 (0.098–0.363) | 0.501 (0.394–0.567) | 0.301 (0.238–0.348) |
| Reduction of 1% in stroke risk for drug B | 0.116 (0.054–0.234) | 0.294 (0.244–0.331) | 0.590 (0.515–0.640) |
| MXL | | | |
| Baseline | 0.177 (0.020–0.389) | 0.412 (0.306–0.490) | 0.412 (0.306–0.490) |
| Increase cost of $50 for drug B | 0.193 (0.025–0.408) | 0.511 (0.377–0.620) | 0.295 (0.215–0.366) |
| Reduction of 1% in stroke risk for drug B | 0.148 (0.013–0.348) | 0.269 (0.199–0.327) | 0.583 (0.445–0.682) |
| MROL | | | |
| Baseline | 0.243 (0.174–0.336) | 0.379 (0.332–0.413) | 0.379 (0.332–0.413) |
| Increase cost of $50 for drug B | 0.273 (0.200–0.372) | 0.486 (0.415–0.542) | 0.242 (0.203–0.279) |
| Reduction of 1% in stroke risk for drug B | 0.204 (0.142–0.289) | 0.263 (0.226–0.298) | 0.533 (0.474–0.581) |

Data are presented as probability (95% confidence interval)

These are the probabilities that each alternative is chosen as best. Baseline refers to the case when all of the attributes for drug A and drug B are set to zero. Each variation in attribute level is simulated one at a time, and only selected variations have been reported

*MNL* multinomial logit, *MROL* mixed rank ordered logit, *MXL* mixed logit

simply given by the ratio of the mean attribute coefficient to the negative of the estimated cost coefficient. Relaxing the assumption that the cost coefficient is fixed can lead to complications: a normally distributed cost coefficient, for example, leads to a distribution for mWTP that has no defined mean since the cost coefficient can now be equal to zero. Researchers therefore often choose a distribution for the cost coefficient that is constrained to be negative to avoid this problem, such as the negative of a log-normal distribution.[14] While this solves the problem of the mWTP distribution not having a defined mean, it can lead to a non-standard distribution whose mean may not be straightforward to calculate.[15] One solution is to approximate the mean using simulation by taking many draws from the distribution of the attribute coefficient and the price coefficient, calculating the ratio for each draw and taking the average of the calculated ratios. If we take a large number of draws, the resulting average should be close to the true mean of the mWTP distribution.

An alternative to estimating the model in the usual way and calculating mWTP as the ratio of parameters is to reformulate the model so that mWTP is estimated directly. This approach, called estimation in WTP space [54], is appealing as it avoids the complications just described. The recent paper by Ben-Akiva et al. [55] covers this estimation approach in detail with illustrative examples. Estimation in WTP space is supported in both Nlogit and Stata, and is possible to implement in Biogeme by exploiting the option for specifying non-linear utility functions.

The mWTP measures can be conditioned on observed choices to obtain individual-level estimates of mWTP; see, for example, Hole [30] and Greene and Hensher [56]. The individual-level mWTP measures can be useful for policy analysis, for example to identify respondents who are likely to benefit particularly highly from a policy improvement.

### 3.4.2 *Predictive Analysis*

A predictive analysis is an extremely flexible post-estimation tool. It is a convenient way to characterize how predicted probabilities change in response to changes in attributes as well as providing a means to simulate interesting scenarios. We illustrate the former here and refer interested readers to Johar et al. [57] for an application involving policy changes; for applications where the impact of the introduction of a new product is investigated, see Fiebig et al. [34] and Ghijben et al. [11], who operationalize procedures outlined in Train [49].

Again using the MNL, MXL and MROL results from Table 1, consider a base case where all attributes have been set to zero. This produces predicted probabilities for each of the estimation methods given in the rows labelled 'baseline' of Table 5. Subsequent rows then show how these predicted probabilities would change in response to two particular changes in the attributes of drug B. The first is a $AUD50 increase making drug B less attractive; the

[14] Log-normal parameter distributions are supported by all of the packages. The negative of the log-normal can easily be implemented by multiplying the price attribute by −1 before entering the model. This is equivalent to specifying the negative of the price coefficient to be log-normally distributed. The sign of the coefficient can easily be reversed post-estimation.

[15] One exception is when both the attribute coefficient and the negative of the price coefficient are log-normally distributed, in which case the distribution of mWTP is also log-normal.

second, that makes drug B more attractive is a 1% point reduction in stroke risk. Each of these changes in drug B is simulated separately and so the appropriate comparison in each case is with reference to the baseline probabilities.

Nlogit, Stata and Biogeme all have built-in routines for conducting predictive analyses. Confidence intervals for the predictions can be generated by taking many draws from the attribute coefficients, generating the predicted probabilities of interest for each draw and calculating the desired percentiles of the generated distributions.

The implications of IIA are clear from the changes in MNL predicted probabilities in comparison with those for MXL and MROL. For example, the response to the cost change implies a dramatic predicted shift away from drug B but, in the case of MNL, the change in predicted probabilities is such that relativities between, say, drug A and no treatment are maintained at the baseline level: (0.501/0.198) = (0.417/0.165) = 2.55. In contrast, the increase in the predicted share of drug A is proportionally larger in the case of MXL and MROL, implying a more realistic substitution pattern. We note that the above analysis differs from the predicted probability analysis presented in Ghijben et al. [11], where more complex policy scenarios are explored, including the introduction of new medications as well as recalibration to market data.

Marginal effects are essentially a simple form of predictive analysis, in which the probabilities in a baseline scenario are compared with the probabilities in an alternative scenario following a marginal increase in a single attribute of one of the alternatives in the model. When viewed this way, marginal effects can be calculated using the same tools as those used to carry out predictive analyses. Elasticities can be calculated in an analogous way, only that in this case we are looking at the percentage change in the probabilities resulting from a 1% increase in an alternative attribute.

It is worth bearing in mind that a potential issue with using DCE data for predictive analysis is that the data do not embody the market equilibrium. Calibrating the ASCs using market data is therefore strongly advisable where such data are available.

### *3.4.3 Welfare Analysis*

A key behavioural outcome of interest to economists is individual and aggregate WTP and willingness to accept monetary amounts in response to policy changes such as a change in single attributes, multiple attributes or the introduction or removal of entire options from the choice set. Indeed, such values are essential in cost-benefit analysis. Such values can be calculated in post-estimation welfare analysis using the compensating variation (CV). In the case of the MNL model, the CV can be expressed as follows:

$$\mathrm{CV} = \frac{1}{\mu}\left[\ln \sum_{j=1}^{J} e^{V_j^0} - \ln \sum_{j=1}^{J} e^{V_j^1}\right] \tag{11}$$

where $V_j^0$ and $V_j^1$ are the values of the utility function, $V$, estimated in the choice model for each choice option $j$ before and after the quality change, respectively, and $J$ is the number of options in the choice set. The log sum terms in Eq. 11 weight the utility associated with each alternative by the probability of selecting that alternative and as such can be interpreted as the expected utility. The CV therefore calculates the change in expected utility before and after the policy change and scales this utility difference by the marginal utility of income, $\mu$, to provide a monetary and therefore cardinal measure of the change in welfare. Often information on income is unavailable, in which case the coefficient on the price attribute (which represents the marginal disutility of price) can be used as the negative of the marginal utility of income. In fact any quantitative numeraire would work—see for example Lancsar et al. [58], who use the marginal utility of a quality-adjusted life-year (QALY) as the numeraire. Calculation of the CV involves harnessing the coefficients estimated in the choice model along with the values of the attributes of interest and can easily be undertaken by hand or in standard software packages (e.g. using nlcom in Stata, which also produces confidence intervals). The interested reader is referred to Lancsar and Savage [58] for further discussion of the theory and methods for such calculations.

## 4 Discussion

Choice modelling is a critical component of undertaking a DCE but to date has received less attention in terms of guidance than other components. As we highlight in Sect. 2, researchers face a number of decisions when analysing DCE data and arriving at a final model. Some decisions are resolved by the choice problem and design (e.g. binary vs. multinomial choice, which variables are independently identified in the experimental design, whether the choice is labelled therefore allowing for the possibility of alternative specific utility functions, or generic, etc.) and others are specification decisions that need to be made on a case-by-case basis (e.g. functional form of specific variables, which could be linear, quadratic, logarithmic, etc., forms of heterogeneity to be explored, etc.). Such decisions are not necessarily linear and sequential; instead, many are simultaneous and interdependent.

There is no single model that we would recommend in all cases. Each have a number of advantages and possible

disadvantages depending on the research question being addressed. In selecting a modelling approach, we would recommend finding a model that addresses the researcher's questions of interest and provides a reasonable device with which to represent the choice at hand. Ultimately, it is still a model and all models involve assumptions. The question to address is which assumptions are most appropriate (or have minimum detriment) to the research questions being explored. It is important to note that the choice of model to be estimated is not only dependent on the research objectives, study design, etc. but is also constrained by what can be estimated given the data a researcher has (including such issues as quantity and quality) and so is based on considerations from both Sects. 2 and 3.

MNL was for decades the workhorse of choice modelling and we recommend it as a natural first model to estimate. Where to go next after MNL is not always clear and depends on the research objectives, but a basic first step would be the estimation of a mixed logit model to account for the panel structure of the data, providing more reliable standard errors and move away from proportional substitution (by relaxing IIA). It also allows for unobserved preference heterogeneity by allowing coefficients to vary randomly across individuals. Whether one takes a Frequentist or Bayesian approach to the estimation of mixed logit in part comes down to preferences of the researcher, but, with the use of simulation methods, the distinction between the two approaches is becoming less pronounced, and recent evidence suggests little difference in estimates [59]. Focus on mixed logit in the health economics literature has often been motivated by interest in unobserved heterogeneity. To our minds, the other two reasons for exploring mixed logit are at least as important.

Having said that, exploration of heterogeneity can be important and has received much attention. If one views the distribution of preference heterogeneity to be discrete rather than continuous, a latent class model would be appropriate. By allowing for different preference parameters between classes, an advantage of latent class modelling is that it allows heterogeneity to be interpreted in terms of class type and class membership. Another form of heterogeneity gaining attention is SH. A modification to the MNL leads to the heteroscedastic logit, which allows for between-person differences in scale to be modelled as a function of covariates. Alternatively, interest in unobserved SH could lead to the SH model. G-MNL offers a very flexible approach that nests several of the standard models discussed in Sect. 2 including mixed logit, heteroscedastic logit and SH models.

When exploring heterogeneity, key decisions to make include which form(s) of heterogeneity are most of interest to the researcher (e.g. preferences or scale, unobserved or observed; noting that these need not be mutually exclusive). While both observed and unobserved heterogeneity can be important, and indeed can be explored within the same model, a distinction between the two is that the latter often improves model fit but is not always readily interpretable. In contrast, observed heterogeneity is interpretable in relation to known covariates (e.g. age, gender, past experience, etc.), thereby potentially generating useful implications for policy and practice. Ultimately, the source(s) of heterogeneity to be explored depends on the research questions, the assumptions researchers are prepared to make and what is revealed by the data. Whichever model is estimated, it is important to be cognizant of the implications of the model (and associated assumptions) chosen for the conclusions that can be drawn.

We also provided model and estimation procedures for best–worst and best–best DCEs, including ROL and SBWMNL. The fact that both models can be estimated by expanding the data as described in Sect. 3 and then applying MNL has a particularly advantageous feature in that it is straightforward to estimate more sophisticated versions of these base models, allowing for non IIA, correlated errors and various forms of heterogeneity. For example, all of the models discussed in Sect. 2 for estimation with first best-data (mixed logit, G-MNL, etc.) can be estimated simply by running such commands on the expanded best–worst or best–best data. Correct data set-up is therefore crucial, and we offer advice on a useful way to check this.

More generally, the advantages and limitations of best–worst data collection have been outlined elsewhere. As Lancsar et al. [14] note, a key advantage is the generation of more data relative to a standard DCE, which can prove particularly useful when sample size constraints exist (due to budget considerations or when the population from which the sample is being drawn is itself small); even when sample size is not a constraint, it can prove an efficient way to generate a given quantity of data or simply provide more data. The additional data can also prove particularly useful for the estimation of models for single individuals [31, 35].

We presented several standard software options in Sect. 3. An advantage of Stata and Nlogit beyond estimation is that they provide comprehensive data management. In contrast, Biogeme requires external data management but is very flexible in estimation; it is also free. Which a researcher selects will in part depend on personal preferences, particularly if they have already invested time and resources in a particular software package. We also offered advice on data set-up and best practice in terms of estimation, including issues such as choice of starting values, number of draws in estimation, etc.

When it comes to interpretation of results, the parameter estimates from choice models are typically not of intrinsic interest and indeed parameters often cannot meaningfully be compared because of the different scales on which they

are measured, some of which may be quantitative (e.g. time, cost, risk, etc.) and others of which may be qualitative (provider type, location, etc.) [33]. It is therefore surprising that many researchers stop after generating attribute coefficients without undertaking post-estimation, particularly given post-estimation is not difficult and provides useful insights. We strongly encourage researchers to harness model parameters in post-estimation analysis to both improve interpretation and to produce measures that are relevant to policy and practice. At a minimum, we suggest the calculation of marginal rates of substitution, but—depending on the goals of the research—additional analysis could include predicted uptake or demand, elasticities and welfare analysis.

As with all methods, validity is crucial. Internal validity has received considerable attention in the health economics literature, including checking signs of estimated parameters accord with a priori expectations and the testing of axioms of consumer theory (e.g. Ryan and Bate [60] and Ryan and San Miguel [61]). Lancsar and Louviere [62] caution against deleting data on the basis of such tests. Lancsar and Swait [9] provide a new and more comprehensive conceptualization of external validity, which advocates that its investigation should be broader than the comparison of final outcomes and predictive performance and indeed encompasses process validity. They suggest innovative ways in which the broader definition can be pursued in practice, starting from the initial conception and design of a DCE through to model and post-estimation. Most relevant to the modelling stage of DCE research is the possible extension of the basic random-utility choice modelling framework in an attempt to more closely replicate reality, for example to account for decision rule selection and choice set formation.

We did not set out to be exhaustive in our coverage of either choice models or software, instead focusing on standard models that can be estimated in standard commonly used software. There are, of course, interesting extensions to these core models that warrant attention for particular research questions, often requiring bespoke coding and estimation. One interesting stream of choice modelling is to account for different underlying decision rules and processing strategies. Two examples of this are choice set formation, championed by Swait and colleagues in the general choice modelling literature (e.g. Swait and Erdem [20]) and starting to be used in health (e.g. Fiebig et al. [63]), and attribute non-attendance (e.g. Hensher and Greene [64], Lagarde [65] and Hole et al. [66]), where the latter can also arise from the broader issue of excessive cognitive burden [67]. Another useful stream of choice modelling is data fusion. We discussed the need to calibrate ASCs for market data where such data are available, particularly for welfare and forecasting analysis. A natural extension is more complete data fusion, where stated-preference data collected in a DCE can be combined with revealed-preference data either from observed choices [68, 69] or indeed from linking or embedding experiments in other data collection (cross sectional, panel, experimental, randomised controlled trials) more generally to harness the advantages of the various data sources [70]. There are, of course, other interesting choice modelling extensions, and we refer the interested reader to the *Handbook of Choice Modelling* [71] for a recent survey of cutting-edge choice models and estimation issues on the research frontier.

## 5 Conclusion

As the use of DCEs and DCE results by researchers, policy makers and practitioners in the health sector continues to increase, so too will the importance placed on the theory and methods underpinning the approach in general and the analysis and interpretation of the generated choice data in particular. As this guide has highlighted, choice of modelling approach depends on a number of factors (research questions, study design and constraints such as quality/quantity of data), and decisions regarding analysis of choice data are often simultaneous and interdependent. When faced with such decisions, we hope the theoretical and practical content of this paper proves useful to researchers not only within but also beyond health economics.

**Acknowledgements** The authors thank Peter Ghijben, Emily Lancsar and Silva Zavarsek for making available the data used in Ghijben et al. [11]. All authors jointly conceived the intent of the paper, drafted the manuscript and approved the final version.

**Compliance with Ethical Standards**

**Funding** No funding was received for the preparation of this paper.

**Conflicts of interest** Emily Lancsar is funded by an ARC Fellowship (DE1411260). Emily Lancsar, Denzil Fiebig and Arne Risa Hole have no conflicts of interest.

**Data Availability Statement** The data and estimation code used in this paper are available in the ESM. Supplementary Appendix 1 contains the data in 'long form', Supplementary Appendix 2 contains the data in 'wide form', Supplementary Appendix 3 contains the data in 'exploded form' and Supplementary Appendix 4 contains the Stata, Nlogit and Biogeme estimation code.

## References

1. Louviere JJ, Hensher DA, Swait JD. Stated choice methods: analysis and applications. Cambridge: Cambridge University Press; 2000.

2. de Bekker-Grob EW, Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. Health Econ. 2012;21(2):145–72.
3. Clark MD, Determann D, Petrou S, Moro D, de Bekker-Grob EW. Discrete choice experiments in health economics: a review of the literature. Pharmacoeconomics. 2014;32(9):883–902.
4. Viney R, Lancsar E, Louviere J. Discrete choice experiments to measure consumer preferences for health and healthcare. Expert Rev Pharmacoeconomics Outcomes Res. 2002;2(4):319–26.
5. Lancsar E, Louviere J. Conducting discrete choice experiments to inform healthcare decision making. Pharmacoeconomics. 2008;26(8):661–77.
6. Bridges J, Hauber A, Marshall D, Lloyd A, Prosser L, Regier D, et al. A checklist for conjoint analysis applications in health: report of the ISPOR Conjoint Analysis Good Research Practices Taskforce. Value Health. 2011;14(4):403–13.
7. Coast J, Horrocks S. Developing attributes and levels for discrete choice experiments using qualitative methods. J Health Serv Res Policy. 2007;12(1):25–30.
8. Johnson FR, Lancsar E, Marshall D, Kilambi V, Mühlbacher A, Regier DA, et al. Constructing experimental designs for discrete-choice experiments: report of the ISPOR conjoint analysis experimental design good research practices task force. Value Health. 2013;16(1):3–13.
9. Lancsar E, Swait J. Reconceptualising the external validity of discrete choice experiments. Pharmacoeconomics. 2014;32(10):951–65.
10. Hauber B, Gonzalez J, Groothuis-Oudshoorn C, Prior T, Marshall D, Cunningham C, et al. Statistical methods for the analysis of discrete choie experiments: a report of the ISPOR conjinta analysis good research practice task force. Value Health. 2016;19:300–15.
11. Ghijben P, Lancsar E, Zavarsek S. Preferences for oral anticoagulants in atrial fibrillation: a best–best discrete choice experiment. Pharmacoeconomics. 2014;32(11):1115–27.
12. Dillman DA. Mail and internet surveys: the tailored design method. New York: Wiley; 2000.
13. Tourangeau R, Rips LJ, Rasinski K. The psychology of survey response. Cambridge: Cambridge University Press; 2000.
14. Lancsar E, Louviere J, Donaldson C, Currie G, Burgess L. Best worst discrete choice experiments in health: methods and an application. Soc Sci Med. 2013;76:74–82.
15. Louviere JJ, Flynn TN, Marley A. Best–worst scaling: theory, methods and applications. Cambridge: Cambridge University Press; 2015.
16. Bartels R, Fiebig DG, van Soest A. Consumers and experts: an econometric analysis of the demand for water heaters. Empir Econ. 2006;31(2):369–91.
17. King MT, Hall J, Lancsar E, Fiebig D, Hossain I, Louviere J, et al. Patient preferences for managing asthma: results from a discrete choice experiment. Health Econ. 2007;16(7):703–17.
18. Jung K, Feldman R, Scanlon D. Where would you go for your next hospitalization? J Health Econ. 2011;30(4):832–41.
19. Harris KM, Keane MP. A model of health plan choice: inferring preferences and perceptions from a combination of revealed preference and attitudinal data. J Econ. 1998;89(1):131–57.
20. Swait J, Erdem T. Brand effects on choice and choice set formation under uncertainty. Market Sci. 2007;26(5):679–97.
21. Swait J, et al. Context dependence and aggregation in disaggregate choice analysis. Market Lett. 2002;13:195–205.
22. McFadden D. Conditional logit analysis of qualitative choice behavior. Berkeley, CA: University of California; 1974.
23. McFadden D. Disaggregate behavioral travel demand's RUM side. A 30 year retrospective. Travel Behav Res. 2000:17–63.
24. Maddala G. Limited dependent and qualitative variables in econometrics. Cambridge: Cambridge University Press; 1983.
25. Fiebig DG, Keane MP, Louviere J, Wasi N. The generalized multinomial logit model: accounting for scale and coefficient heterogeneity. Market Sci. 2010;29(3):393–421.
26. Hess S, Rose JM. Can scale and coefficient heterogeneity be separated in random coefficients models? Transportation. 2012;36(6):1225–39.
27. Revelt D, Train K. Mixed logit with repeated choices: households' choices of appliance efficiency level. Rev Econ Stat. 1998;80(4):647–57.
28. Brownstone D, Train K. Forecasting new product penetration with flexible substitution patterns. J Econ. 1998;89(1):109–29.
29. Hall J, Fiebig DG, King MT, Hossain I, Louviere JJ. What influences participation in genetic carrier testing? Results from a discrete choice experiment. J Health Econ. 2006;25(3):520–37.
30. Hole AR. Modelling heterogeneity in patients' preferences for the attributes of a general practitioner appointment. J Health Econ. 2008;27(4):1078–94.
31. Louviere JJ, Street D, Burgess L, Wasi N, Islam T, Marley AA. Modeling the choices of individual decision-makers by combining efficient choice experiment designs with extra preference information. J Choice Model. 2008;1(1):128–64.
32. Keane M, Wasi N. Comparing alternative models of heterogeneity in consumer choice behavior. J Appl Econ. 2013;28(6):1018–45.
33. Lancsar E, Louviere J, Flynn T. Several methods to investigate relative attribute impact in stated preference experiments. Soc Sci Med. 2007;64(8):1738–53.
34. Fiebig DG, Knox S, Viney R, Haas M, Street DJ. Preferences for new and existing contraceptive products. Health Econ. 2011;20(S1):35–52.
35. Lancsar E, Louviere J. Estimating individual level discrete choice models and welfare measures using best–worst choice experiments and sequential best–worst MNL. Sydney: University of Technology, Centre for the Study of Choice (Censoc); 2008. p. 08-004.
36. Scarpa R, Notaro S, Louviere J, Raffaelli R. Exploring scale effects of best/worst rank ordered choice data to estimate benefits of tourism in alpine grazing commons. Am J Agric Econ. 2011;93(3):813–28.
37. Punj GN, Staelin R. The choice process for graduate business schools. J Market Res. 1978;15(4):588–98.
38. Chapman RG, Staelin R. Exploiting rank ordered choice set data within the stochastic utility model. J Market Res. 1982;19(3):288–301.
39. Beggs S, Cardell S, Hausman J. Assessing the potential demand for electric cars. J Econ. 1981;17(1):1–19.
40. Gu Y, Hole AR, Knox S. Fitting the generalized multinomial logit model in Stata. Stata J. 2013;13(2):382–97.
41. Hole AR. Fitting mixed logit models by using maxium simulated likelihood. Stata J. 2007;7:388–401.
42. Pacifico D, Yoo HI. lclogit: a stata module for estimating latent class conditional logit models via the Expectation-Maximization algorithm. Stata J. 2013;13(3):625–39.
43. Baker MJ. Adaptive Markov chain Monte Carlo sampling and estimation in Mata. Stata J. 2014;14(3):623–61.
44. Suits DB. Dummy variables: mechanics v. interpretation. Rev Econ Stat. 1984;66:177–80.
45. Bech M, Gyrd-Hansen D. Effects coding in discrete choice experiments. Health Econ. 2005;14(10):1079–83.
46. Hole AR, Yoo I. The use of heuristic optimization algorithms to facilitate maximum simulated likelihood estimation of random parameter logit models. J R Stat Soc C. 2017;. doi:10.1111/rssc.12209.
47. Czajkowski M, Budziński W. An insight into the numerical simulation bias—a comparison of efficiency and performance of different types of quasi Monte Carlo simulation methods under a

wide range of experimental conditions. In: Environmental Choice Modelling Conference; Copenhagen; 2015.
48. Bhat CR. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. Transp Res Part B: Methodol. 2001;35(7):677–93.
49. Train KE. Discrete choice methods with simulation. Cambridge: Cambridge University Press; 2009.
50. Hess S, Train KE, Polak JW. On the use of a Modified Latin Hypercube Sampling (MLHS) method in the estimation of a Mixed Logit Model for vehicle choice. Transp Res Part B: Methodol. 2006;40(2):147–63.
51. Garrido RA. Estimation performance of low discrepancy sequences in stated preferences. In: 10th International Conference on Travel Behaviour Research; Lucerne; 2003.
52. Munger D, L'Ecuyer P, Bastin F, Cirillo C, Tuffin B. Estimation of the mixed logit likelihood function by randomized quasi-Monte Carlo. Transp Res Part B: Methodol. 2012;46(2):305–20.
53. Hole AR. A comparison of approaches to estimating confidence intervals for willingness to pay measures. Health Econ. 2007;16(8):827–40.
54. Train K, Weeks M. Discrete choice models in preference space and willingness-to-pay space: Applications of simulation methods in environmental and resource economics. Berlin: Springer; 2005. p. 1–16.
55. Ben-Akiva M, McFadden D, Train K. Foundations of stated preference elicitation consumer behavior and choice-based conjoint analysis. In: 2015, Society for economic measurement annual conference, Paris, 24 July 2015.
56. Greene WH, Hensher DA. A latent class model for discrete choice analysis: contrasts with mixed logit. Transp Res Part B: Methodol. 2003;37(8):681–98.
57. Johar M, Fiebig DG, Haas M, Viney R. Using repeated choice experiments to evaluate the impact of policy changes on cervical screening. Appl Econ. 2013;45(14):1845–55.
58. Lancsar E, Wildman J, Donaldson C, Ryan M, Baker R. Deriving distributional weights for QALYs through discrete choice experiments. J Health Econ. 2011;30:466–78.
59. Lancsar E, Savage E. Deriving welfare measures from discrete choice experiments: inconsistency between current methods and random utility and welfare theory. Health Econ. 2004;13(9):901–7.
60. Elshiewy O, Zenetti G, Boztug Y. Differences between classical and Bayesian estimates for mixed logit models: a replication study. J Appl Econ. 2017;32(2):470–76.
61. Ryan M, Bate A. Testing the assumptions of rationality, continuity and symmetry when apply- ing discrete choice experiments in health care. Appl Econ Lett. 2001;8:59–63.
62. Ryan M, San MF. Revisiting the axiom of completeness in health care. Health Econ. 2003;12:295–307.
63. Lancsar E, Louviere J. Deleting, "irrational" responses from discrete choice experiments: a case of investigating or imposing preferences? Health Econ. 2006;15(8):797–811.
64. Fiebig DG, Viney R, Knox S, Haas M, Street DJ, Hole AR, et al. Consideration sets and their role in modelling doctor recommendations about contraceptives. Health Econ. 2017;26(1):54–73.
65. Hensher DA, Greene WH. Non-attendance and dual processing of common-metric attributes in choice analysis: a latent class specification. Empir Econ. 2010;39(2):413–26.
66. Lagarde M. Investigating attribute non-attendance and its consequences in choice experiments with latent class models. Health Econ. 2013;22(5):554–67.
67. Hole AR, Kolstad JR, Gyrd-Hansen D. Inferred vs. Stated attribute non-attendance in choice experiments: a study of doctors' prescription behaviour. J Econ Behav Organ. 2013;96:21–31.
68. Flynn TN, Bilger M, Finkelstein EA. Are efficient designs used in discrete choice experiments too difficult for some respondents? A case study eliciting preferences for end-of-life care. Phamacoeconomics. 2016;34(3):273–84.
69. Mark TL, Swait J. Using stated preference and revealed preference modeling to evaluate prescribing decisions. Health Econ. 2004;13(6):563–73.
70. Ben-Akiva M, Bradley M, Morikawa TJ, Benjamin T, Novak H, Oppewal H, et al. Combining revealed and stated preferences data. Market Lett. 1994;5(4):335–49.
71. Lancsar E, Burge P. Choice modelling research in health economics. In: Hess S, Daly A, editors. Handbook of choice modelling. Cheltenham, UK: Edward Elgar Press; 2014. p. 675–87.
72. Hess S, Daly A. Handbook of choice modelling. Cheltenham: Edward Elgar Publishing; 2014.