VALUE IN HEALTH 19 (2016) 300–315

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

FEATURED ARTICLES
ISPOR Task Force Report

Statistical Methods for the Analysis of Discrete Choice Experiments: A Report of the ISPOR Conjoint Analysis Good Research Practices Task Force

A. Brett Hauber, PhD[1,], Juan Marcos González, PhD[1], Catharina G.M. Groothuis-Oudshoorn, PhD[2], Thomas Prior, BA[3], Deborah A. Marshall, PhD[4], Charles Cunningham, PhD[5], Maarten J. IJzerman, PhD[2], John F.P. Bridges, PhD[6]*

[1]*RTI Health Solutions, Research Triangle Park, NC, USA;* [2]*Department of Health Technology and Services Research, University of Twente, Enschede, The Netherlands;* [3]*Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA;* [4]*Department of Community Health Sciences, Faculty of Medicine, University of Calgary and O'Brien Institute for Public Health, Calgary, Alberta, Canada;* [5]*Department of Psychiatry and Behavioural Neuroscience, Michael G. DeGroote School of Medicine, McMaster University, Hamilton, Ontario, Canada;* [6]*Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA*

ABSTRACT

Conjoint analysis is a stated-preference survey method that can be used to elicit responses that reveal preferences, priorities, and the relative importance of individual features associated with health care interventions or services. Conjoint analysis methods, particularly discrete choice experiments (DCEs), have been increasingly used to quantify preferences of patients, caregivers, physicians, and other stakeholders. Recent consensus-based guidance on good research practices, including two recent task force reports from the International Society for Pharmacoeconomics and Outcomes Research, has aided in improving the quality of conjoint analyses and DCEs in outcomes research. Nevertheless, uncertainty regarding good research practices for the statistical analysis of data from DCEs persists. There are multiple methods for analyzing DCE data. Understanding the characteristics and appropriate use of different analysis methods is critical to conducting a well-designed DCE study. This report will assist researchers in evaluating and selecting among alternative approaches to conducting statistical analysis of DCE data. We first present a simplistic DCE example and a simple method for using the resulting data. We then present a pedagogical example of a DCE and one of the most common approaches to analyzing data from such a question format—conditional logit. We then describe some common alternative methods for analyzing these data and the strengths and weaknesses of each alternative. We present the ESTIMATE checklist, which includes a list of questions to consider when justifying the choice of analysis method, describing the analysis, and interpreting the results.

Keywords: conjoint analysis, discrete choice experiment, stated-preference methods, statistical analysis.

Introduction

Since the middle of the 1990s, there has been a rapid increase in the use of conjoint analysis to measure the preferences of patients and other stakeholders in health applications [3–8]. Although early applications were used to quantify process utility [9,10], more recent applications have focused on patient preferences for health status [11,12], screening [13], prevention [14,15], pharmaceutical treatment [16,17], therapeutic devices [18,19], diagnostic testing [20,21], and end-of-life care [22,23]. In addition, conjoint analysis methods have been used to study decision making among stakeholders other than patients, including clinicians [24–26], caregivers [25,27], and the general public [28,29].

"Conjoint analysis" is a broad term that can be used to describe a range of stated-preference methods that have respondents rate, rank, or choose from among a set of experimentally controlled profiles consisting of multiple attributes with varying levels. The most common type of conjoint analysis used in health economics, outcomes research, and health services research (hereafter referred to collectively as "outcomes research") is the discrete choice experiment (DCE) [6,8]. The premise of a DCE is that choices among sets of alternative profiles

Conflicts of interest: All authors represent the ISPOR Conjoint Analysis Statistical Analysis Good Research Practices Task Force.

* *Address correspondence to:* A. Brett Hauber, RTI Health Solutions, 200 Park Offices Drive, Research Triangle Park, NC 27709.
E-mail: abhauber@rti.org.

1098-3015$36.00 – see front matter

http://dx.doi.org/10.1016/j.jval.2016.04.004

Background to the Task Force

The International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Conjoint Analysis Statistical Analysis Good Research Practices Task Force is the third ISPOR Conjoint Analysis Task Force. It builds on two previous task force reports, "Conjoint Analysis Applications in Health—A Checklist: A Report of the ISPOR Good Research Practices for Conjoint Analysis Task Force" [1] and the "ISPOR Conjoint Analysis Experimental Design Task Force" [2]. The Conjoint Analysis Checklist report developed a 10-point checklist for conjoint analysis. The checklist items were as follows: 1) the research question, 2) the attributes and levels, 3) the format of the question, 4) the experimental design, 5) the preference elicitation, 6) the design of the instrument, 7) the data collection plan, 8) the statistical analysis, 9) the results and conclusions, and 10) the study's presentation.

This first task force determined that several items, including experimental design (checklist item 4) and methods for analyzing data from conjoint analysis studies (checklist item 8), deserved more detailed attention. Thus, the ISPOR Conjoint Analysis Experimental Design Task Force focused on experimental design to assist researchers in evaluating alternative approaches to this difficult and important element of a successful conjoint analysis study.

This third task force report describes a number of commonly used options available to researchers to analyze data generated from studies using a particular type of conjoint analysis—the discrete choice experiment—and the types of results generated by each method. This report also describes the issues researchers should consider when evaluating each analysis method and factors to consider when choosing a method for statistical analysis.

The Conjoint Analysis Statistical Analysis Good Research Practices Task Force proposal was submitted to the ISPOR Health Science Policy Council for evaluation in December 2012. The council recommended the proposal to the ISPOR Board of Directors, and it was subsequently approved in January 2013.

Researchers with experience in stated preferences and discrete choice experiments working in academia and research organizations in Canada, the Netherlands, and the United States were invited to join the task force's leadership group. The leadership group met via regular teleconferences to identify and discuss present analytical techniques, develop the topics and outline, and prepare drafts of the manuscript. A list of leadership group members (coauthors) is also available on the task force's Web page (http://www.ispor.org/Conjoint-Analysis-Statistical-Methods-Guidelines.asp).

All task force members, as well as primary reviewers, reviewed many drafts of the report and provided frequent feedback as both oral and written comments. Preliminary findings and recommendations were presented twice in forum presentations at ISPOR Montreal (2014) and ISPOR Milan (2015). Comments received during these presentations were addressed in subsequent drafts of the report.

In addition, the draft task force report was sent to the ISPOR Preference-Based Methods Review Group twice. All comments were considered, and most were substantive and constructive. The comments were discussed by the task force and addressed as appropriate in revised drafts of the report. Once consensus was reached by all task force members, the final report was submitted to *Value in Health* in March 2016.

are motivated by differences in the levels of the attributes that define the profiles. By controlling the attribute levels experimentally and asking respondents to make choices among sets of profiles in a series of choice questions, a DCE allows researchers to effectively reverse engineer choice to quantify the impact of changes in attribute levels on choice. The estimates of these impacts reflect the strength of preference for changes in attribute levels. We refer to these estimates of strength of preference, which are sometimes called "part-worth utilities," as "preference weights." This task force report focuses on motivating and reviewing the most common statistical methods that are presently used in outcomes research to analyze data from a DCE.

Most applications of a DCE in outcomes research are designed to test hypotheses regarding strength of preference for, relative importance of, and trade-offs among attributes that define the research question. A DCE is appropriate for this type of research because preference weights derived from a DCE are estimated on a common scale and can be used to calculate ratios describing the trade-offs respondents are willing to make among the attributes. Examples of these trade-offs include estimates of money equivalence (willingness to pay) [30,31], risk equivalence (maximum acceptable risk) [18,32], or time equivalence [33,34] for various changes in attributes or attribute levels. Although the underlying premise and the mechanics of using a DCE are similar in market research and outcomes research, the objectives of using a DCE typically differ between these disciplines. In DCEs conducted in market research or marketing science (which are often referred to as "choice-based conjoint analysis studies"), the objective is often to predict choices with as much precision as possible. In contrast, DCEs in outcomes research tend to focus more on understanding preferences for changes in attributes and attribute levels. Therefore, most statistical analyses of DCEs in outcomes research are geared toward estimating interpretable preference weights rather than toward estimating the model with the greatest predictive power.

The application of DCEs to measuring preferences for health and health care has benefited from a growing literature on methods [35–39], including the two previous conjoint analysis task force reports from the International Society for Pharmacoeconomics and Outcomes Research (ISPOR). This present report builds on the first of these reports, "Conjoint Analysis Applications in Health—A Checklist: A Report of the ISPOR Good Research Practices for Conjoint Analysis Task Force" [1]. The checklist outlines the steps to take for the development, analysis, and publication of conjoint analyses. The checklist items are as follows: 1) research question, 2) attributes and levels, 3) construction of tasks, 4) experimental design, 5) preference elicitation, 6) instrument design, 7) data collection plan, 8) statistical analyses, 9) results and conclusions, and 10) study presentation. This task force report is designed to provide researchers practicing in outcomes research with a better understanding of the methods commonly used to analyze data from DCEs in this field. It focuses exclusively on item 8 in the checklist (statistical analysis). Therefore, the examples used in this report are designed specifically to provide the context for describing the properties of alternative analysis methods and are not meant to reflect or imply good research practices for survey design or development or presentation of study results.

Understanding the characteristics and appropriate analysis of preference data generated by DCE surveys is critical to conducting a well-designed DCE. Good research practices for the statistical analysis of DCE data involve understanding the characteristics of alternative methods and ensuring that interpretation of the results is accurate, taking into account both the assumptions made during the course of the analysis and the strengths and limitations of the analysis method. Despite the

growing use of conjoint analysis methods in outcomes research, there remains inconsistency in the statistical methods used to analyze data from DCEs [1,4,5]. Given this inconsistency, the task force agreed that good research practices in the analysis of DCE data must start with ensuring that researchers have a good understanding of the fundamentals of DCE data and the range of statistical analysis methods commonly used in applications of DCEs in outcomes research. Although there are several other key methodological references that may be useful to more experienced researchers [40–42], the task force realized that these texts may be unfamiliar to a more general audience of researchers. The task force determined that a pragmatic introduction to different statistical analysis methods, highlighting differences among methods and identifying the strengths and limitations of each method, was needed.

This report starts with the basic idea behind deriving preference weights from DCE data by describing two simple approaches to calculating preference weights from a simplistic DCE with a very limited number of attributes and levels. The purpose of providing these examples is to help readers understand some of the basic properties of choice data. We then present a slightly more complex, but still relatively simple, pedagogical example—a three-attribute, two-alternative, forced-choice DCE. Using this example, we describe alternative approaches to coding the data generated by this type of DCE and describe one possible method for constructing the data set. We then describe the analysis of data from this example using a conditional logit model consistent with the random utility model of choice [35,40,43]. Because most of the other commonly used methods for analyzing DCE data are, in effect, variations on the conditional logit model, we then describe extensions of conditional logit that can be used to analyze the same DCE data. These extensions include random-parameters (mixed) logit (RPL), hierarchical Bayes (HB), and latent-class finite-mixture model (LCM) analysis methods. To demonstrate the differences in the properties of each of these analysis methods, we present the results of each method as applied to a common simulated data set. The report concludes with a summary of the strengths and limitations of each method described in the report and provides the ESTIMATE checklist, which has a series of questions to consider when justifying the choice of analysis method, describing the analysis, and interpreting the results.

A Simplistic Example

To understand the basic concepts underlying the analysis of DCE data, consider a simplistic example described using the case of a hypothetical, over-the-counter analgesic. If we assume that the relevant analgesics can be described by three attributes, each with two possible levels (time to onset of action can be 30 minutes or 5 minutes; duration of pain relief can be 4 hours or 8 hours; and formulation can be tablets or capsules), then these can be combined into eight possible profiles. Although these eight profiles can be combined into a large number of distinct pairs, we could use a main-effects orthogonal design [2] to generate an experimental design consisting of four paired-profile choice tasks (Appendix Table 1).

If there were three respondents who completed all four choice tasks, we would have 12 observations with which to estimate a model. Assume the respondents answered as follows: respondent 1 (B, B, A, B), respondent 2 (A, B, A, A), and respondent 3 (A, B, B, A). The simplest way to analyze these choices is to count how many times each level of each attribute was chosen by counting the number of times each attribute level was chosen by each respondent, summing these totals across all respondents, and dividing this sum by the number of times each attribute level was presented across the three respondents to calculate a score for each attribute level (Appendix Table 2). Although there appears to be some heterogeneity in preferences among the respondents (e.g., respondent 1 appears to prefer capsules to tablets, whereas respondents 2 and 3 appear to prefer tablets to capsules), we can still infer sample-level preferences from these data. Across the sample, a 5-minute onset was preferred to a 30-minute onset, a 4-hour duration was preferred to an 8-hour duration, and tablets were preferred to capsules.

We can also use regression analysis to linearly relate the probability of choosing one profile over another to all the characteristics of the profiles simultaneously. This model assumes that the probability of choosing one profile is a linear function of the attribute levels in the profile. Thus, the model can be described as follows:

$$\Pr(\text{choice}) = \beta_0 + \sum\nolimits_i \beta_i X_i, \quad (1)$$

where X_i is the level of attribute i, β_0 is the intercept, and β_i is the preference weight for attribute i.

This model relates choice to the attribute levels in each profile. Other versions of the linear probability model relate choice to differences in attribute levels between two profiles. We do not present the linear probability model based on attribute-level differences in this report.

The linear probability model defines a relationship between choices and attribute levels that can be leveraged to estimate preference weights through various linear regression models. One such estimator is ordinary least squares (OLS). With OLS, the estimates of the intercept (β_0) and the preference weights (β_i) are defined as the set of values of the estimates that minimizes the difference between the observations in the data and the estimated model. This difference is known as the "sum of the squared residuals."

To set up the data for the regression analysis in this example, let Onset = 1 if the onset of action is 30 minutes and Onset = 0 if it is 5 minutes in profile A, Duration = 1 if the duration is 4 hours and Duration = 0 if it is 8 hours in profile A, and Tablet = 1 if the formulation is a tablet and Tablet = 0 if the formulation is a capsule in profile A. Finally, let Choice = 1 if profile A was chosen and Choice = 0 if Profile B was chosen.

Using OLS on the binary choice variable Y (Y = 1 if profile is chosen and Y = 0 if the profile is not chosen), the model takes the following form:

$$Y = \alpha + \beta_1 \text{Onset} + \beta_2 \text{Duration} + \beta_3 \text{Tablet} + \varepsilon, \quad (2)$$

where α is the intercept, and ε is the random error term, and the conditional expectation of the binary variable Y equals the probability that profile A is chosen. Then the estimated linear probability function is as follows:

$$\Pr(\text{choice}) = \Pr(Y=1) = 0.33 - 0.33\text{Onset} + 0.33\text{Duration} + 0.33\text{Tablet}. \quad (3)$$

The coefficients from this simplistic model can be interpreted as marginal probabilities: an analgesic with a 30-minute onset of action is 33% less likely to be chosen than one with a 5-minute onset of action; an analgesic with a 4-hour duration of effect is 33% more likely to be chosen than one with an 8-hour duration of effect; and an analgesic in the form of a tablet is 33% more likely to be chosen than one in the form of a capsule. From this we can infer that, on average, respondents in the sample prefer faster onset of action, shorter duration of effect, and tablets over capsules. Comparing the results of this regression with the results of the count analysis, we find that the regression coefficient on each attribute is simply the difference between the sample-level scores for the levels of that attribute. Specifically, the score for a 30-minute onset of action was 0.33 less than the score for a 5-minute onset of action, the score for a 4-hour duration of effect was 0.33 higher than the score for an 8-hour duration of effect, and the score for tablet form was 0.33 higher than the score for capsule form. It is important to note that the preference information in the linear probability model is perfectly

confounded with the probability of choice associated with changes in attribute levels. That is, the measure of how much more a respondent prefers a change in the level of one attribute is the marginal change in the probability of choice for alternatives that differ only in that attribute.

OLS yields unbiased and consistent coefficients and has the advantage of being easy to estimate and interpret. Nevertheless, OLS is subject to a number of limitations when used to analyze DCE data. When using OLS, researchers must assume that the errors with which they measure choices are independent and identically distributed with mean 0 and constant variance [44]. In reality, the variance in DCE data changes across choice tasks. In addition, even with estimators other than OLS, linear probability models can produce choice probabilities that are greater than 1 or less than 0 for certain combinations of attribute levels. For these reasons, among others, linear probability methods are rarely used to analyze DCE data.

A Pedagogical Example

To describe the more common alternatives for analyzing data from a DCE, we provide a pedagogical example of a DCE. As mentioned earlier, we define a DCE in which each choice task presents a pair of alternatives. Respondents are asked to choose between the two profiles in each pair. Each profile is defined by three attributes, and each attribute has three possible levels. Thus, this example is a three-attribute, two-alternative, forced-choice experiment. In this example, each profile is a medication alternative. The three attributes that define each medication alternative are efficacy, a side effect, and a mode of administration. Table 1 presents the attributes and levels used to create the profiles in this example. For simplicity, we assume that efficacy is measured on a numeric scale from 0 to 10, with higher values representing better outcomes. Efficacy is a variable with numeric, and thus naturally ordered, levels. The levels of the side effect are severities (mild, moderate, or severe). The side effect levels are not only categorical but also naturally ordered from less severe to more severe. The levels of the mode of administration are daily tablets, weekly subcutaneous injections, and monthly intravenous infusions. The levels for mode of administration are categorical, and the ordering of preferences for these levels is unknown a priori. The attributes and levels included in this example are meant to demonstrate the mechanics of generating data from the DCE. Detailed descriptions of the selection of attributes, the determination of the levels and range of levels for each attribute, and the methods used for presenting the profiles and the attribute levels included in each profile are beyond the scope of this report.

Table 1 – Attributes and attribute levels in the pedagogical example.

Attribute	Level
A1: Efficacy	L1: 10 (best level) L2: 5 (middle level) L3: 3 (worst level)
A2: Side effect	L1: Mild L2: Moderate L3: Severe
A3: Mode of administration	L1: One tablet once a day L2: Subcutaneous injection once a week L3: Intravenous infusion once a month

Variable Coding in the Pedagogical Example

One way to think about the data generated by a DCE is that each row in the data set corresponds to a single profile in the DCE. Defining the levels in each row of data can be accomplished in multiple ways. Attributes with numeric levels (e.g., survival time, risk, and cost) can be specified as continuous variables. Using this approach, the level value of the attribute in the profile will appear in the appropriate place in the row. In the pedagogical example, only efficacy can be logically coded as a continuous variable because the levels of both the side effect and the mode of administration are descriptive and thus categorical.

Two commonly used methods for categorical coding of attribute levels are effects coding and dummy-variable coding [41,45]. In each of these coding approaches, one level of each attribute must be omitted. In both effects coding and dummy-variable coding, each nonomitted attribute level is assigned a value of 1 when that level is present in the corresponding profile and 0 when another nonomitted level is present in the corresponding profile. The difference between the two coding methods is related to the coding of the nonomitted levels when the omitted level is present in the profile. With effects coding, all nonomitted levels are coded as –1 when the omitted level is present. With dummy-variable coding, all nonomitted levels are coded as 0 when the omitted level is present.

The coefficient on the omitted level of an effects-coded variable can be recovered as the negative sum of the coefficients on the nonomitted levels of that attribute. Therefore, effects coding yields a unique coefficient for each attribute level included in the study. Each effects-coded coefficient, however, is estimated relative to the mean attribute effect; therefore, statistical tests of significance for each coefficient are not direct tests of the statistical significance of differences between estimated coefficients on two different levels of the same attribute. With dummy-variable coding, each coefficient estimated by the model is a measure of the strength of preference of that level relative to the omitted level of that attribute. Statistical tests of significance for each coefficient reflect the statistical significance of the difference between that preference weight and the omitted category. Effects coding and dummy-variable coding yield the same estimates of differences in preference weights between attribute levels [46]. For this reason, in most cases, the decision to use effects coding or dummy-variable coding of variables should be based on ease of interpretation of the estimates from the model, and not on the expectation that one type of coding will provide more information than the other. See Bech and Gyrd-Hansen [45] for further discussion of differences between effects coding and dummy-variable coding.

Data Generated by the Pedagogical Example

One possible way to set up the data generated by this DCE is to construct two rows of data for each choice task for each respondent, one row for each alternative. If each respondent is presented with 10 choice tasks generated by an appropriate experimental design and there are 200 respondents, the total data set will have 4000 rows. In the pedagogical example with two alternatives, the first row of data for each choice task will include the attribute levels that appear in the first profile in the pair presented in that choice task. The second row of data for the same choice task will include the attribute levels for the second profile in that pair. In addition, each row of data will include a choice dummy variable equal to 1 if the chosen profile for the choice task corresponds to that row of data or 0 otherwise.

Table 2 presents an example of this type of data setup for a two-alternative choice task using the attributes and levels from the pedagogical example in which the first medicine alternative (medicine A) has an efficacy of 10, has a severe side effect, and is

Table 2 – Example data setup for the pedagogical example using effects coding.

ID no.*	Task	Choice	Efficacy		Side effect		Mode of administration	
			10 (best level) (L1)	5 (middle level) (L2)	Mild (L1)	Moderate (L2)	One tablet once a day (L1)	Subcutaneous injection once a week (L2)
1	1	0	1	0	–1	–1	0	1
1	1	1	0	1	1	0	–1	–1

* The ID number is the respondent number.

administered by subcutaneous injection once a week and the second medicine alternative (medicine B) has an efficacy of 5, has a mild side effect, and is administered by an intravenous infusion once a month, using effects coding for all attributes and assuming that level 3 is the omitted level for each attribute.

The ID number in the first column of Table 2 is the respondent number. Task is the number of the choice task in the series. Choice indicates that the first respondent chose medicine B when presented with the choice task as described earlier. The level of the side effect in medicine A is severe (the omitted level). Therefore, the nonomitted levels of this attribute are coded as –1 in the first line of data. In addition, the level for the mode of administration in medicine B is the omitted level for this attribute and the nonomitted levels for this attribute are both coded as –1 in the two columns farthest to the right in the second line of data. If the levels in this example were dummy-variable–coded, then the values for the nonomitted levels of the side effect for medicine A and the values for the nonomitted levels of the mode of administration for medicine B would all be set to 0.

Conditional Logit

Choice data from a two-alternative, forced-choice DCE as described in the pedagogical example are most often analyzed using a limited dependent-variable model because the left-hand-side variable in the regression is typically a 1 for the alternative that was chosen in a given choice task or a 0 for the alternative that was not chosen in that choice task. The basic limited dependent-variable method used to analyze data generated by this type of experiment is conditional logit. Conditional logit relates the probability of choice among two or more alternatives to the characteristics of the attribute levels defining those alternatives. In a DCE, the elements describing the alternatives are the attribute levels used to define each profile in the choice task. Conditional logit was shown by McFadden [43] to be consistent with random utility theory. The novelty in McFadden's use of the logit model is that he applied this model to choice behavior that was consistent with economic theory and derived a regression model that relates choices to the characteristics of the alternatives available to decision makers. McFadden used the term "conditional logit" to describe this innovation. McFadden originally applied this framework to observed transportation choices. His work laid the foundation for what is now known as conjoint analysis [40] involving hypothetical or stated choices.

Using random utility theory, the utility associated with an alternative or profile is assumed to be a function of observed characteristics (attribute levels) and unobserved characteristics of the alternative. This theoretic framework also assumes that each individual, when faced with a choice between two or more alternatives, will choose the alternative that maximizes his or her utility. The utility function is specified as an indirect utility function defined by the attribute levels in the alternative plus a random error term reflecting the researcher's inability to perfectly measure utility:

$$U_i = V(\beta, X_i) + \varepsilon_i, \tag{4}$$

where V is a function defined by the attribute levels for alternative i, ε_i is a random error term, X_i is a vector of attribute levels defining alternative i, and β is a vector of estimated coefficients. Each estimated coefficient is a preference weight and represents the relative contribution of the attribute level to the utility that respondents assign to an alternative. In conditional logit, ε_i is assumed to follow an independently and identically distributed type 1 extreme-value distribution [43].

The assumption of the extreme-value distribution of ε_i results in a logit model:

$$\Pr(\text{choice} = i) = \frac{e^{V(\beta,\, x_i)}}{\sum_j e^{V(\beta, x_j)}}, \tag{5}$$

where $V(\beta, x_i)$ is the observed portion of the function for alternative i, and i is one alternative among a set of j alternatives. Simply stated, the probability of choosing alternative i is a function of both the attribute levels of alternative i and the attribute levels of all other profiles presented in a choice task. In the case of the two-alternative, forced-choice DCE, there are two alternatives in each choice task, and so $j = 2$. The probability of choosing one profile from the set of two alternatives is 1 minus the probability of choosing the other profile in that choice task. Therefore, neither alternative in the choice task has a choice probability of less than 0% or greater than 100%. In addition, this specification implies that the closer the probability of choosing an alternative in a two-alternative choice task is to 50%, the more sensitive the probability of choosing that alternative is to changes in the attribute levels that define the alternative.

Multinomial logit is similar to conditional logit in that it also can be used to model choices. Both multinomial logit and conditional logit rely on the same statistical assumptions about the relationship between choice and the variables used to explain choice. Nevertheless, multinomial logit is often used to describe models that relate choices to the characteristics of the respondents making the choices, whereas conditional logit relates choices to the elements defining the alternatives among which respondents choose. Although the terms "multinomial logit" and "conditional logit" are sometimes used interchangeably in the literature, we will use the term "conditional logit" in this report because the objective of the DCE is to relate choice to the attribute levels used to define each profile in the choice task.

Results of the Conditional Logit Model Using Effects-Coding and Dummy-Variable Coding

Table 3 presents a set of example results from conditional logit model regressions for the pedagogical example with each of these types of variable coding. In a conditional logit model, a coefficient (or preference weight) and a corresponding standard error are estimated for all but one level of each attribute. A t value

Table 3 – Example results from a conditional logit model.

Attribute	Level	Effects coding				Dummy-variable coding			
		Coefficient	SE	t value	P value	Coefficient	SE	t value	P value
Efficacy	L1	0.26	0.0105	24.89	<0.01	0.55	0.0183	29.85	<0.01
	L2	0.02	0.0104	2.24	0.03	0.31	0.0181	17.03	<0.01
	L3 (omitted category)	−0.28	0.0105	−27.03	<0.01	Constrained to be 0			
Side effect	L1	0.32	0.0105	30.18	<0.01	0.66	0.0184	35.71	<0.01
	L2	0.02	0.0103	2.02	0.04	0.36	0.0181	19.91	<0.01
	L3 (omitted category)	−0.34	0.0106	−32.07	<0.01	Constrained to be 0			
Mode of administration	L1	0.03	0.0104	2.49	0.01	0.24	0.0182	13.08	<0.01
	L2	0.18	0.0105	17.61	<0.01	0.40	0.0181	21.67	<0.01
	L3 (omitted category)	−0.21	0.0105	−20.04	<0.01	Constrained to be 0			
Log likelihood			−17,388				−17,388		
Log likelihood of model without predictors			−18,715				−18,715		
AIC			34,788				34,788		
BIC			34,841				34,841		

AIC, Akaike information criterion; BIC, Bayesian information criterion; SE, standard error.

and a *P* value are often calculated for each estimated preference weight. With effects coding, each *P* value is a measure of the statistical significance of the difference between the estimated preference weight and the mean effect of the attribute. With dummy-variable coding, each *P* value is a measure of the statistical significance of the difference between the estimated preference weight and the omitted category. In some statistical packages, confidence intervals at the 95% confidence level are also provided for each preference weight. Table 3 also includes Akaike information criterion (AIC), Bayesian information criterion (BIC), and log-likelihood values for each model; all measures of model fit are discussed later in this report. The variance-covariance matrices for the effects-coded and dummy-variable–coded conditional logit models are presented in Appendix Tables 3 and 4, respectively.

As noted earlier, the preference weight (coefficient) on the omitted level of an effects-coded variable is recovered by calculating the negative of the sum of the estimated preference weights for all nonomitted levels of the attribute. For example, in the results in Table 3, the estimated preference weight for the omitted level of efficacy in the model with effects-coded attribute levels is equal to −0.28 (= −[0.26 + 0.02]). The standard error of the omitted category can be calculated by using simulation methods or by using the following simple formula:

$$\text{Standard error of } L3 = \sqrt{\text{Variance of } L1 + \text{Variance of } L2 + 2 \times \text{Covariance between } L1 \text{ and } L2}. \quad (6)$$

Interpreting the Results of the Conditional Logit Model: Preference Weights

The preference weights for the effects-coded and dummy-variable–coded conditional logit models are presented in Figures 1

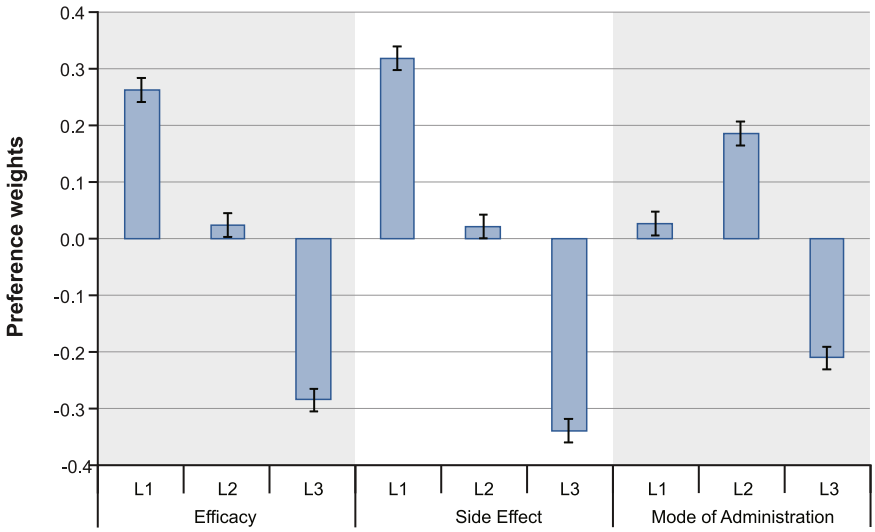

Fig. 1 – Preference weights for the conditional logit model (effects-coded)

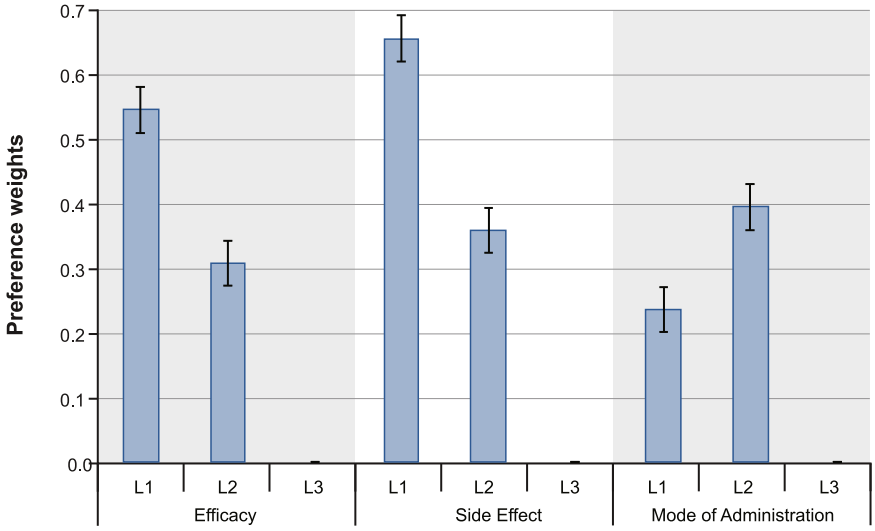

Fig. 2 – Preference weights for the conditional logit model (dummy-coded)

and 2, respectively. Both figures clearly indicate that higher levels of efficacy are preferred to (have higher preference weights than) lower levels of efficacy, a less severe side effect is preferred to a more severe side effect, and subcutaneous injections are preferred to one tablet once a day, which, in turn, is preferred to an intravenous infusion once a month. The presentation of the results in Figures 1 and 2 is intended for pedagogical purposes only. An endorsement of good research practices for the presentation of results from a DCE survey is beyond the scope of the present task force report.

Figures 1 and 2 highlight that the absolute values of the preference weights and the confidence intervals about the mean preference weights differ between the effects-coded and the dummy-coded models but that differences in the preference weights among levels within each attribute are the same in both models. For example, the preference weight for the highest level of efficacy (L1) is 0.55 in the dummy-coded model and 0.26 in the effects-coded model. Nevertheless, the effect of increasing efficacy from L2 to L1 is 0.24 (= 0.55 – 0.31) in the dummy-coded model and 0.24 (= 0.26 – 0.02) in the effects-coded model. The effect of reducing side effects from L2 to L1 is 0.30 (= 0.66 – 0.36) in the dummy-coded model and 0.30 (= 0.32 – 0.02) in the effects-coded model. The similarities between the effects-coded and the dummy-coded models highlight an important aspect of the conditional logit model. That is, the absolute values of preference weights alone have no meaningful interpretation. Preference weights measure relative preference, which means that only changes between attribute-level estimates and the relative sizes of those changes across attributes have meaningful interpretations. In the effects-coded and the dummy-coded models that we have used in the examples, directly comparing the parameter estimates for the efficacy attribute in each model would have erroneously suggested that the two models provided different results or different information on the relative impact that these attribute levels have on choice. As the figures demonstrate, differences between the two model specifications are merely due to differences in variable coding, and no additional information is obtained from either model. This implies that estimates from conditional logit models cannot be compared across models directly and require close attention to the relationships between levels of different attributes. Results from models that share the same specification can be compared using a test proposed by Swait and Louviere [47].

The difference in preference weights between the best or most preferred level of an attribute and the worst or least preferred level of the same attribute provides an estimate of the relative importance of that attribute over the range of levels included in the experiment. For example, the relative importance of a 7-point change in efficacy from L1 (a value of 10) to L3 (a value of 3) is 0.54 (0.26 – [–0.28] in the effects-coded model and 0.55 – 0 in the dummy-coded model). The only difference between these estimates of relative importance is due to rounding. The value of a change in side effect severity from severe to mild is 0.66 (0.32 – [–0.34] in the effects-coded model and 0.66 – 0 in the dummy-coded model). Therefore, reducing side effect severity from severe to mild yields approximately 1.2 (0.66 ÷ 0.54) times as much utility as increasing efficacy by 7 points (from 3 to 10). Likewise, changing the mode of administration from an intravenous infusion once a month to a subcutaneous injection once a week yields only 0.72 (0.39 ÷ 0.54) times as much utility as increasing efficacy by 7 points.

These results can also be used to estimate the rate at which respondents would be willing to trade among the attributes in the experiment. For example, a reduction in side effect severity from moderate to mild yields an increase of 0.30 in utility. The reduction in efficacy (from the highest level of 10) that exactly offsets this increase in utility would be approximately 5.4 points. This change in efficacy is calculated as the 0.24 reduction in utility achieved by moving from an efficacy value of 10 to an efficacy value of 5, plus the change in efficacy between an efficacy value of 5 and an efficacy value of 3 that would yield a reduction of 0.06 in utility. The necessary additional change in utility is equal to one-fifth of the difference (0.06 ÷ 0.30) between an efficacy value of 5 and an efficacy value of 3. Interpolating between these two points and adding the value to the change from an efficacy value of 10 to an efficacy value of 5 results in a total change in efficacy from 10 to 4.6 necessary to offset the increase in utility from reducing the severity of the side effect.

Goodness of Fit

Conditional logit models do not allow for the calculation of an *R*-squared measure of absolute goodness of fit with which many readers may be familiar. Instead, several measures that mimic an *R*-squared calculation have been developed using the log likelihood of these models. Log likelihood is an indicator of the relative explanatory power of a model. Higher (less negative) values in the log likelihood are associated with a greater ability of a model to explain the pattern of choices in the data. Table 3

presents two log-likelihood values for each model. In each model, one log-likelihood value is for the estimated model and the other corresponds to a model that includes only a constant for all but one of the alternatives. Although log-likelihood values alone cannot be used as a measure of model fit because they are a function of sample size, they can be used to calculate goodness-of-fit measures such as the likelihood ratio chi-square test and McFadden's pseudo *R*-squared, which are commonly reported in software packages. The likelihood ratio chi-square test provides a way to determine whether including attribute-level variables significantly improves the model fit compared with a model without any attribute-level variables and indicates whether one or more of the preference weights are expected to be different from 0. Models with a higher likelihood ratio can be assumed to fit the data better. The critical value for this test can be calculated as follows:

$$\text{Likelihood ratio } \chi^2 = -2(\text{LL model without predictors} - \text{LL of model}), \quad (7)$$

where LL is log-likelihood value. The likelihood ratio has a chi-squared distribution with degrees of freedom equal to the number of preference weights estimated in the model minus the number of alternative-specific constants included in the model (no alternative-specific constants are included in the model in the pedagogical example).

Although many software packages provide the likelihood ratio test statistic (Equation 7), there are also ways to test for improvements in fit between an unrestricted model (that estimates all preference weights) and a restricted model (that forces one or more preference weights to be 0). In that case, the critical value for this test can be calculated as follows:

$$\text{Likelihood ratio } \chi^2 = -2(\text{LL of restricted model} - \text{LL of unrestricted model}), \quad (8)$$

with degrees of freedom equal to the number of restrictions (preference weight estimates forced to be 0). For example, an unrestricted model may be one that estimates preference weights for all attributes in a DCE, whereas the restricted model forces the weights for one of the attributes to be 0 (in essence eliminating the attribute from the analysis). In such a case, the likelihood ratio chi-square test statistic indicates whether estimating the weights for the eliminated attribute improves model fit.

McFadden's pseudo *R*-squared is calculated using the following formula:

$$\text{McFadden's pseudo } R^2 = 1 - \frac{\text{LL of model}}{\text{LL model without predictors}}. \quad (9)$$

McFadden's pseudo *R*-squared can be 0 if all the preference weights on attribute levels are constrained to be 0, but the measure can never reach 1. The measure improves with the number of explanatory variables, which is not an issue when comparing models with the same number of explanatory variables as in the models with effects- and dummy-coded variables, but can be problematic when comparing models that have different numbers of explanatory variables or different specifications of the utility function. Note that McFadden's pseudo *R*-squared for the effects-coded model [0.0709 (= 1 – [–17,388 ÷ –18,715])] is identical to that for the dummy-coded model [0.0709 (= 1 – [–17,388 ÷ –18,715])], indicating that both models explain choices equally well. Although McFadden's pseudo *R*-squared provides a measure of relative (rather than absolute) model fit, a measure from 0.2 to 0.4 can be considered a good model fit [48].

In cases in which two models have different explanatory variables, other measures of model fit are commonly provided by statistical packages. The most common of these measures are the AIC and the BIC. Both criteria are specified as $-2LL + K\gamma$, where LL is log-likelihood value of the full model, *K* is the number of parameter estimates corresponding to the number of explanatory variables in the model, and γ is a penalty constant that changes between AIC ($\gamma = 2$) and BIC ($\gamma = \ln[\text{sample size}]$). The premise of the calculations behind these measures is different from that of the calculation of pseudo *R*-squared measures. AIC and BIC evaluate the plausibility of the models focusing on minimizing information loss rather than evaluate improvements in the adequacy of the model to explain responses in the data as done in the pseudo *R*-squared measures. Also, contrary to pseudo *R*-squared measures, models with lower AIC and BIC measures are preferred over models with higher measures. As with McFadden's pseudo *R*-squared measures, AIC and BIC measures are identical across the models in Table 3.

The Conditional Logit Model with a Continuous Variable

An alternative specification of conditional logit as applied to the pedagogical example is to code the efficacy attribute as a linear continuous variable with the efficacy values shown (3, 5, and 10) in each alternative (Appendix Table 5). The estimated preference weight for efficacy is an estimate of the relative marginal utility of a 1-unit change in the efficacy outcome. All changes of more than 1 unit in the efficacy are assumed to be proportional to a 1-unit change. Given this specification, the effect of a 1-unit change in efficacy on utility is assumed to be constant over the range of levels of this attribute included in the experiment. For example, the utility change resulting from a 9-unit change in the efficacy measure is assumed to be 9 times the marginal utility of a 1-unit change in the efficacy measure.

Two pieces of evidence suggest that a linear continuous specification for efficacy may be inappropriate given the data in this experiment. First, results from the categorical models suggest that the relative marginal utility of a 1-unit change in the efficacy measure is more important when the change in efficacy is between 3 (level 3) and 5 (level 2) than when the change in efficacy is between 5 (level 2) and 10 (level 1). Specifically, the difference in the relative preference weights between levels 1 and 2 of the efficacy attribute is 0.24 (implying a 0.05 change in preference weight associated with a 1-unit change in the level of efficacy over this range of levels), and the difference in the relative preference weights between levels 2 and 3 of the efficacy attribute is 0.30 (implying a 0.15 change in preference weight associated with a 1-unit change in the level of efficacy over this range of levels).

The second indication that the linear continuous specification of efficacy may be incorrect is the difference in the goodness of fit between the models in which the efficacy attribute is modeled as a linear continuous variable and the models in which the efficacy attribute is modeled as a categorical variable. This change in model fit is evident if we calculate McFadden's pseudo *R*-squared for the model with the continuous specification, which is 0.0686 (= 1 – [–17,432 ÷ –18,715]) and compare it with the same measure for either of the categorical specifications. In addition, the increase in the AIC and BIC calculations for the linear continuous model suggests that modeling efficacy as categorical provides a better fit to the data than modeling efficacy as linear and continuous. The reduction in the model goodness of fit suggests that the proportionality assumption imposed on the marginal effect of the efficacy attribute does not fit the data as well as the categorical preference weights.

These results should not be interpreted to mean that a continuous specification of numeric attribute levels is incorrect. Instead, these results suggest that a researcher should not assume that a linear continuous specification is appropriate for attributes with numeric levels. Instead, researchers should test to see whether a linear specification of a continuous variable is appropriate by examining the results of a model in which the

numeric levels are treated as categorical. If a linear continuous specification is not correct, the researcher can test for the appropriateness of other nonlinear functional forms of this variable.

Limitations of Conditional Logit

There are two fundamental limitations of the conditional logit model. First, the model assumes that choice questions measure utility equally well (or equally poorly) across all respondents and choice tasks (scale heterogeneity). Second, conditional logit does not account for unobserved systematic differences in preferences across respondents (preference heterogeneity). Although historically these two problems have been considered distinct issues of the conditional logit model, more recent discussions acknowledge that the two problems are related [49].

The first problem is scale heterogeneity [50]. The conditional logit model assumes that the ratio of estimated preference weights is exactly the same if both the numerator and the denominator are multiplied by the same value [51]. Using conditional logit, this value is inversely related to a constant model variance normalized to 1 for all respondents. If, however, the variance (i.e., the ability of the model to estimate utility) is not constant across all respondents or across all choice tasks, differences in model estimates between different respondents or across different choice tasks may appear different when in fact they are not. For example, assume that the overall importance (difference between the highest and the lowest preference weights) of efficacy and the side effect are 0.80 and 0.20, respectively, for one group of choice tasks or one group of respondents in the sample and that the same values are 1.60 and 0.40, respectively, for another group of choice tasks or respondents. The absolute values of the relative importance estimates for these two attributes are twice as large for the second group as they are for the first group; nevertheless, the ratio of the relative importance of efficacy to the side effect is the same in both groups. The scale in this case is constant. In some cases, however, assuming that the scale is constant across respondents ignores potential systematic variations in the model variance across choice questions or groups of individuals in the sample and can result in biased estimates of preference weights [52].

The second potential problem with conditional logit is preference heterogeneity—potential differences in relative preferences across respondents in the sample. In effect, the conditional logit model assumes that all respondents have the same preferences that will yield a single set of preference weights. The conditional logit model does not account for systematic variations in preferences across respondents. Failing to account for heterogeneity in preferences can lead to biased estimates of the preference weights.

The relationship between the two potential problems is clearer if we acknowledge that preference heterogeneity can be correlated across attribute levels. That is, preferences for multiple attributes can increase or decrease together among specific individuals in the sample. If variations in preferences are allowed to be correlated across respondents, preference heterogeneity can behave exactly as scale heterogeneity does. That is, if a respondent's preferences move proportionately together across attributes, the ratios of preferences for any pair of attributes remain unchanged, whereas the absolute values of the preference weights would be different by a constant scaling factor, just as they would with scale heterogeneity.

Alternatives to Conditional Logit

Several alternative statistical techniques have been proposed to overcome some, but not all, of the limitations of conditional logit. Specifically, these models are able to address conditional logit's inability to account for correlation among multiple responses from each individual or heterogeneity in preferences across the sample. Alternative techniques continue to be developed; this report, however, focuses on three statistical methods that are commonly used in health applications of DCEs. These statistical methods include RPL, HB, and LCM analyses. All three of these methods are based on limited dependent-variable models consistent with random utility theory and ensure that the predicted probability of choosing any alternative is between 0% and 100%. Each of these techniques is more statistically complex than conditional logit, but each has potential advantages. These three approaches are discussed in the following sections. Knowing that effects-coded and dummy-coded models provide the same information, we will use the results from only the effects-coded models to characterize the alternatives to conditional logit in the following sections of this report.

Random-Parameters Logit

Like conditional logit, RPL (also called "mixed-logit") is a method that assumes that the probability of choosing a profile from a set of alternatives is a function of the attribute levels that characterize the alternatives and a random error term that adjusts for individual-specific variations in preferences. Unlike conditional logit that estimates only a set of coefficients capturing the mean preference weights of the attribute levels, RPL yields both a mean effect and a standard deviation of effects across the sample. That is, RPL explicitly assumes that there is a distribution of preference weights across the sample reflecting differences in preferences among respondents, and it models the parameters of that distribution for each attribute level. The choice probability of the RPL model is as follows:

$$\Pr(\text{choice}_n = i) = \frac{e^{V(\tilde{\beta}_n, x_i)}}{\sum_j e^{V(\tilde{\beta}_n, x_j)}}, \tag{10}$$

where n indexes respondents in the sample, $\tilde{\beta}_n = f(\beta, \sigma | \upsilon_n)$, β and σ are parameters to be estimated on the basis of systematic variations in preferences across individuals in the sample given the variable υ_n characterizing individual-specific heterogeneity, and $f(\cdot)$ is a function determining the distribution of $\tilde{\beta}_n$ across respondents, given parameters β and σ. Commonly, $\tilde{\beta}$ is assumed to be normally distributed with mean β and standard deviation σ, which means the following:

$$\begin{aligned}\tilde{\beta}_n &= f(\beta, \sigma | \upsilon_n) \\ &= \beta + \sigma \upsilon_n, \text{ for } \upsilon_n \sim N(0, 1) \text{ across respondents in the sample.}\end{aligned} \tag{11}$$

With this assumption about $f(*)$, the choice probability of conditional logit is a special case of the RPL choice probability where $\sigma = 0$. Larger (smaller) standard deviations indicate greater (lesser) variability in preference weights across respondents. Nevertheless, little direct guidance is available to determine the appropriate functional form for the distribution of preferences across respondents. Because individual preference weights are not directly interpretable, it is often difficult to determine the distributional characteristics of preferences in any given sample a priori.

Table 4 presents example results for an RPL model on choice data from the pedagogical example assuming that each preference weight is normally distributed across the sample. The results in Table 4 contain mean preference weight estimates, plus estimates of standard deviations for all estimated preference weight. The mean preference weight estimates are for effects-coded variables and are interpreted in the same way in which the results of the conditional logit model (Table 3) are interpreted; coefficients from the two models, however, should not be expected to be the same numerically. Mean preference weights

Table 4 – Example results from a random-parameters logit model.

Attribute	Level	Coefficient	SE	t value	P value
	Mean estimates				
Efficacy	L1	0.32	0.0153	20.84	<0.01
	L2	0.03	0.0124	2.32	0.02
	L3 (omitted category)	–0.35	0.0162	–21.45	<0.01
Side effect	L1	0.39	0.0162	23.84	<0.01
	L2	0.02	0.0137	1.67	0.10
	L3 (omitted category)	–0.41	0.0181	–22.54	<0.01
Mode of administration	L1	0.03	0.0116	2.52	0.01
	L2	0.24	0.0187	12.56	<0.01
	L3 (omitted category)	–0.26	0.0190	–13.94	<0.01
	Standard deviation estimates				
Efficacy	L1	0.32	0.0160	20.05	<0.01
	L2	0.16	0.0198	8.35	<0.01
	L3 (omitted category)	0.49	0.0254	19.16	<0.01
Side effect	L1	0.35	0.0161	21.75	<0.01
	L2	0.25	0.0167	15.19	<0.01
	L3 (omitted category)	0.60	0.0218	27.76	<0.01
Mode of administration	L1	0.10	0.0265	3.67	<0.01
	L2	0.47	0.0169	27.88	<0.01
	L3 (omitted category)	0.57	0.0290	19.66	<0.01
Log likelihood of model	–16,657				
Log likelihood of model without predictors	–18,715				
AIC	33,337				
BIC	33,436				

AIC, Akaike information criterion; BIC, Bayesian information criterion; SE, standard error.

from RPL are typically larger in magnitude than the preference weights from the same model estimated using conditional logit, but the relative relationships (e.g., ratios of changes in attributes) among preference weights for different attributes are typically similar between the two sets of coefficients. Each standard deviation indicates the distribution about the corresponding mean preference weight.

One issue with RPL models is that the maximum simulated likelihood estimation used to fit RPL models relies on a simulation technique that can produce different answers if the parameters of the simulations are not set to be the same across regressions. These parameters include the simulation random seed, the number of draws, and the type of draws taken. Researchers should estimate the model with multiple starting points to ensure that the model converges to a stable solution. The estimator sometimes requires a lengthy estimation process that is not guaranteed to converge under certain circumstances and also requires specifying the number of simulations to be used in a regression; ensuring that these numbers are adequate adds burden to the researcher. The model is increasingly available in various software packages, but not all model features are available in all packages.

Hierarchical Bayes

Conditional logit and RPL are used to estimate preferences over a population. Conditional logit yields estimates of mean preference weights for a sample. RPL yields estimates of both mean preference weights and the expected distribution of preference weights across the sample. Although RPL results take into account individual-specific effects, it models these effects as deviations from the mean population parameters. In contrast, HB models reverse the problem to generate preference estimates for each individual in the sample and only supplement these individual-specific estimates with aggregate preference information to the degree that individual-specific preference information is insufficient.

Like RPL, the underlying choice-probability model in HB is conditional logit [53,54]. This choice-probability model, however, is used to model responses from each individual, and not all observations in the sample. With HB, individual results are used to construct the (joint posterior) distribution of preference weights across respondents, including the mean and standard deviation for the preference weight for each attribute level. As with RPL, to calculate the mean and standard deviation of preferences for attribute levels, HB requires that the researcher assume the form of the distribution of each preference weight.

HB models evaluate choices in two levels:

1. The likelihood level, or lower level, in which individual choices are modeled using conditional logit:

$$\Pr(\text{choice}=i)=\frac{e^{V(\beta_n,x_i)}}{\sum_j e^{V(\beta_n,x_j)}}. \quad (12)$$

2. The sample level, or upper level, which is often assumed to be multivariate-normal (or normal for each preference weight), although other distributions, including lognormal, triangular, and uniform, have been used. This level characterizes the variation of preferences across respondents:

$$\beta_n \sim N(b,W). \quad (13)$$

An algorithm (Gibbs sampler) estimates iteratively the lower level and the upper level, individual-specific preference weight parameters (β_n), the overall preference mean (b), and the variance-covariance matrix of preferences across respondents (W). Essentially, b is calculated as the sample average of the β_n, and W is calculated as the sample variance-covariance of the β_n. Each β_n tends toward a value that maximizes the likelihood the model explains the pattern of responses to the series of choice tasks multiplied by the upper-level function characterizing the

Table 5 – Example results comparing RPL and HB models.

Attribute	Level	Mean		Standard deviation estimates	
		RPL	HB	RPL	HB
Efficacy	L1	0.32	0.35	0.32	0.43
	L2	0.03	0.03	0.16	0.33
	L3	−0.35	−0.38	0.49	0.76
Side effect	L1	0.39	0.43	0.35	0.50
	L2	0.02	0.03	0.25	0.42
	L3	−0.41	−0.46	0.60	0.92
Mode of administration	L1	0.03	0.03	0.10	0.36
	L2	0.24	0.26	0.47	0.66
	L3	−0.26	−0.29	0.57	1.02

HB, hierarchical Bayes; RPL, random-parameters logit.

distribution of preferences with mean b and variance-covariance W. The multiplication of the likelihood value for the respondent by the function of the upper-level distribution essentially keeps estimates of individual-level preference weights consistent with the assumed distribution of the sample preferences and not overly far from the sample preference mean [55].

Table 5 presents results from HB estimated using data from the pedagogical example assuming effects-coded variables and normally distributed variations in preferences across respondents. Table 5 also compares the HB estimates of means and standard deviations of the estimated preference weights with the corresponding results from the RPL model estimated using the same assumptions. Although these results should not be compared numerically, Table 5 shows that both methods produce similar results in this example.

HB estimation can be slower for certain specifications of the distributions of preference weights (e.g., truncated distributions such as uniform and triangular distributions). HB estimation is implemented in fewer software packages than is RPL. Understanding the raw output of the HB estimation method requires some knowledge of sampling methods and Bayesian statistics. In some software packages, however, aggregate results are included in addition to the individual-specific estimates. These aggregate results could include sample-level means and standard deviations.

The means of the preference weight distributions in HB are similar to the mean preference weights estimated using conditional logit or RPL [56,57]. The standard deviation for each preference weight represents how different that preference weight is across respondents. Larger standard deviations mean that respondents differ in their perception of that attribute level. If the objective of the research is to estimate preference weights for each individual in the sample, or when the sample size is small, Bayesian procedures may provide a better approach to analysis than the RPL model because the inference using the joint posterior distribution of preference weights can be conducted for any sample size [57]. Also, the HB procedure does not require the assumption of a common scale across respondents imposed in the conditional logit and RPL models.

Latent-Class Finite-Mixture Model

The LCM assumes that attributes of the alternatives can have heterogeneous effects on choices across a finite number of groups or classes of respondents. To account for this heterogeneity, the model assumes that there are classes or segments within a sample such that each class has preference weights that are identical within the class and that are systematically different from preference weights in other classes. Within each class, the preference weights are estimated using conditional logit. The number of classes is prespecified by the researcher. For example, if the analyst chooses to model the sample using three classes, three different sets of conditional logit coefficients are estimated, one for each class.

In the LCM, the choice probability is defined as follows:

$$\Pr(\text{choice}=1)=\sum\nolimits_q \Pr\left(\text{choice}=\text{i}|\beta_q\right)\pi_q, \tag{14}$$

where π_q is a class-probability function indicating the probability of being in each of the different classes. The class-probability function is a specific multinomial logit function that can include only a constant term or can include explanatory variables relating the probability of class membership to respondent characteristics. The probabilities of class membership must sum to 1, so the class-probability functions of all but one class are identifiable. The class-probability function for the omitted class does not change freely. As a consequence, the explanatory variables in the class-probability function must be interpreted as those respondent characteristics that increase or decrease the probability of being in one class relative to the probability of being in the omitted class.

The choice probability within a class q is estimated using conditional logit:

$$\Pr\left(\text{choice}=\text{i}|\beta_q\right)=\frac{e^{V(\beta_q,x_i)}}{\sum_j e^{V(\beta_q,x_j)}}. \tag{15}$$

Table 6 presents example results for an LCM with two classes using the effects-coded data from the pedagogical example. It includes a set of conditional logit model coefficients for each of the two classes and a class-probability function indicating the probability that a respondent is in class 1. The class-probability function in this example includes only a constant and does not include any explanatory variables to test for the influence of individual characteristics on the probability of class membership. The class-probability estimate in this example indicates that, on average, a respondent in the sample has a 40% chance of being in class 1. Because the class-probability function is estimated using multinomial logit and the log-odds coefficient for the constant in the class-probability function is –0.40, the probability of being in class 1 is calculated as follows:

$$40\%=\frac{\exp(-0.40)}{1+\exp(-0.40)}. \tag{16}$$

The results for each class in an LCM can be interpreted as preference weights estimated with conditional logit. Nevertheless, although the preference weights for the classes are estimated within a single model, parameters across classes are generally not directly comparable because they are confounded with a scale parameter that may differ between classes. Therefore, comparisons of preference weights across classes must be

Table 6 – Example results of a latent-class model with two classes.

Attribute	Level	Coefficient	SE	t value	P value
	Class 1				
Efficacy	L1	0.29	0.0246	11.70	<0.01
	L2	0.04	0.0206	2.01	0.04
	L3 (omitted category)	–0.33	0.0236	–13.98	<0.01
Side effect	L1	0.39	0.0295	13.28	<0.01
	L2	–0.02	0.0235	–1.06	0.29
	L3 (omitted category)	–0.37	0.0248	–14.77	<0.01
Mode of administration	L1	0.21	0.0220	9.38	<0.01
	L2	–0.33	0.0355	–9.25	<0.01
	L3 (omitted category)	0.12	0.0276	4.42	<0.01
	Class 2				
Efficacy	L1	0.27	0.0190	14.37	<0.01
	L2	0.01	0.0164	0.80	0.42
	L3 (omitted category)	–0.29	0.0187	–15.34	<0.01
Side effect	L1	0.30	0.0214	14.20	<0.01
	L2	0.05	0.0177	3.09	<0.01
	L3 (omitted category)	–0.36	0.0192	–18.69	<0.01
Mode of administration	L1	–0.09	0.0172	–5.40	<0.01
	L2	0.54	0.0277	19.60	<0.01
	L3 (omitted category)	–0.45	0.0216	–20.84	<0.01
	Class-probability function				
	Constant	–0.40	0.0321	–12.46	<0.01
Log likelihood of model	–16,985				
Log likelihood of model without predictors	–18,715				
AIC	33,996				
BIC	34,103				

AIC, Akaike information criterion; BIC, Bayesian information criterion; SE, standard error.

done by evaluating the ratios of changes in preference weights across attributes. Sometimes, however, it is possible to evaluate the estimates across classes qualitatively to identify relevant differences across classes. For example, the utility difference between L1 and L2 in the mode of administration is positive (0.55) for class 1; for class 2, the same difference is negative (–0.63). Thus, one tablet once a day is preferred to a subcutaneous injection once a week in class 1, whereas the opposite is true for class 2.

In the LCM, there is no control for the effect of multiple observations for each respondent. Usually it is assumed that, given the class assignment, multiple observations from the same respondent are independent. Moreover, just as with conditional logit and RPL, LCM does not control for potential scale heterogeneity within each class; some software packages, however, include scale adjustments. As with RPL, researchers using LCM should estimate the model with multiple starting points to ensure that the model converges to a stable solution. Finally, determining the appropriate number of classes to include in a model is challenging. AIC or BIC can be used to provide an objective determination of the correct number of classes. Nevertheless, using AIC may risk overfitting the model by having too many classes and using BIC may risk underfitting the model by having too few classes [58]. Therefore, when determining the appropriate number of classes in an LCM, the researcher must consider the number of classes required to address the underlying research questions and the ease of interpretation of multiple classes when the number of classes is large.

Discussion

This report describes a number of commonly used limited dependent-variable analysis methods to analyze DCE data. Many researchers may be uncertain as to which method is the most appropriate in any given circumstance, and, to a certain extent, there is no clear consensus on which methods are the best. The first step in selecting a method is to understand the properties of DCE data and the properties of the alternative methods available to analyze it. Therefore, this report aimed to provide a description of these methods and the advantages and limitations of each. The advantages and limitations are summarized in Appendix Table 6. We then provide a checklist of questions to consider when justifying the choice of analysis method, describing the analysis, and interpreting the results (Table 7).

Conditional logit is the first limited dependent-variable analysis method described in this report. Although conditional logit has a number of limitations, understanding conditional logit is fundamental to understanding the other analysis methods described in this report because these other analysis methods are, in effect, extensions of conditional logit. Conditional logit can be used to estimate average preferences across a sample but does not account for preference heterogeneity. Conditional logit is available in many software packages often used to analyze DCE data. In addition, it is commonly used in LCMs or as a first step in RPL or HB analysis. Because conditional logit does not account for the panel nature of the data, the results of a conditional logit analysis could be biased. In addition, as a maximum-likelihood method, conditional logit models may not converge, especially for small samples or in the presence of substantial preference heterogeneity. The model, however, requires the smallest samples for convergence among the options described in this report.

RPL is an analysis method that accounts for the panel nature of DCE data and allows preferences to vary across respondents in the sample. It is becoming more commonly available in statistical software packages, but it is relatively more difficult to use and requires the researcher to make assumptions about which random parameters to include in the analysis and the distributions of those random parameters. Because RPL estimates both

Table 7 – The ESTIMATE checklist.

ESTIMATE	Recommendation
Estimates	Describe the choice of parameter *estimates* resulting from the model appropriately and completely, including • Whether each variable corresponds to an effects-coded level, a dummy-coded level, or a continuous change in levels • Whether each variable corresponds to a main effect or interaction effect • Whether continuous variables are linear or have an alternative functional form
Stochastic	Describe the *stochastic* properties of the analysis, including • The statistical distributions of parameter estimates • The distribution of parameter estimates across the sample (preference heterogeneity) • The variance of the estimation function, including systematic differences in variance across observations (scale heterogeneity)
Trade-offs	Describe the *trade-offs* that can be inferred from the model, including • The magnitude and direction of the attribute-level coefficients • The relative importance of each attribute over the range of levels included in the experiment • The rate at which respondents are willing to trade off among the attributes (marginal rate of substitution)
Interpretation	Provide *interpretation* of the results taking into account the properties of the statistical model, including • Conclusions that can be drawn directly from the results • Applicability of the sample, including subgroups or segments, to the population of interest • Limitations of the results
Method	Describe the reasons for selecting the statistical analysis *method* used in the analysis, including • Why the method is appropriate for analyzing the data generated by the experiment • Why the method is appropriate for addressing the underlying research question • Why the method was selected over alternative methods
Assumptions	Describe the *assumptions* of the model and the implications of the assumptions for interpreting the results, including • Assumptions about the error distribution • Assumptions about the independence of observations • Assumptions about the functional form of the value function
Transparent	Describe the study in a sufficiently *transparent* way to warrant replication, including descriptions of • The data setup, including handling missing data • The estimation function, including the value function and the statistical analysis method • The software used for estimation
Evaluation	Provide an *evaluation* of the appropriateness of the statistical analysis method to answering the research question, including • The goodness of fit of the model • Sensitivity analysis of the model specification • Consistency of results estimated using different methods

mean preference weights and distributions for preference weights, it may require larger sample sizes.

HB analysis provides a different way of modeling preference heterogeneity that can allow for the estimation of a choice model for each respondent. Under most circumstances, HB models converge more quickly than models using alternative methods. Also, because HB models estimate preference weights for each respondent in the sample, they fully account for potential issues of preference and scale heterogeneity (each respondent can have different relative preferences, and the absolute values of the preference weights can vary freely across respondents). Nevertheless, because HB uses methods for updating preference estimates iteratively, it may be difficult to describe this method, which may lead to concerns that it is less transparent than other analysis methods. Also, because HB models do not estimate the sample-wide mean preference weights, but rather construct them from the estimated individual preferences, practitioners need to make inferences about overall effects using standard deviations (variation of preferences across respondents) as opposed to using a standard error around a global estimate of preferences. At the same time, it is not possible to test for the significance of preferences for any given individual in the sample.

LCM provides a more parsimonious method for measuring preference heterogeneity and modeling latent classes but often requires specialized software or advanced macros in more standard packages. It requires the researcher to specify the number of classes to be included in the model because objective criteria may result in too few or too many classes. This requires significant judgment on the part of the researcher, which may be difficult to explain to users of the results. In addition, larger sample sizes may be needed to accommodate the increase in the number of parameters to be estimated as the number of classes increases.

Good Research Principles for Statistical Analysis

This report provides a description and an evaluation of many statistical methods used to analyze DCE data on the basis of the consensus of the task force that good research practices for the statistical analysis of DCE data involve understanding the characteristics of alternative methods and ensuring that the interpretation of results is accurate. Taking into account both the assumptions made during the course of the analysis and the strengths and limitations of the analysis method to assist researchers in using the information presented in this report, the task force developed the ESTIMATE checklist composed of questions to consider when justifying the choice of analysis method, describing the analysis, and interpreting the results. Table 7 presents the ESTIMATE checklist questions.

Study Limitations

Even at the inception of this task force, it was clear that this report could not be all things to all readers. Aspects of this report were criticized by certain reviewers for being too simplistic, not giving strong enough guidance on good research practices, or duplicating material that is addressed in other contexts. Consistent with the other ISPOR conjoint analysis task force reports, we have written this report to be most useful to a more general audience that may not have the experience or training in econometrics or choice modeling that some of our reviewers have. The members of the task force were unanimous in the decision to structure this report as a pragmatic introduction to different statistical analysis methods and to provide a checklist of questions to consider when justifying the selection of an analysis method, describing the analysis, and interpreting the results. For this reason, some of the earlier criticisms of this report may still hold. We acknowledge several additional limitations of this report.

One potential limitation is the report's narrow scope. Specifically, the report was developed to focus on describing the properties of alternative analysis methods and did not address other issues related to good research practices for conducting conjoint analysis studies that were identified in the original checklist. There are many aspects of study design that can affect the quality of data generated by a DCE and influence the analysis required to understand the data. The first of these involves the setup of the choice experiment, including the choice of attributes and levels to be included in the DCE and the format of the choice task. The second design issue involves the way in which attributes and attribute levels are described and presented in the choice tasks. Although good survey development practices are critical to generating useful data from a DCE, this report does not address these issues nor the influence that survey development decisions may have on the data generated by the DCE.

This report does not provide a detailed treatment of the traditional theoretical underpinnings or the alternative theoretical considerations emerging in the choice modeling literature. Although the task force members acknowledge the importance of theory and theoretical considerations, we agreed that the primary objective of this report was to provide information to researchers interested in using conjoint analysis as an empirical tool. We believed that providing support for this type of experimental research was of vital importance given that diverse stakeholders, including patient groups and regulators, are increasingly turning to stated-preference research to inform decision making. Conjoint analysis emerged from and benefited from developments in different disciplines, including, but not limited to, economics, psychology, decision science, and marketing science, and researchers from different disciplines may view conjoint analysis through different lenses. The task force did not want potential disagreements regarding theoretical interpretation of conjoint analysis to undermine researchers' understanding of important considerations in the statistical analyses of DCE data.

Perhaps the biggest limitation of this report is that it focuses on a simple pedagogical example—a three-attribute, two-alternative, forced-choice DCE. The task force chose this example because it allowed us to describe some of the fundamental issues involved in the analysis of DCE data while keeping the scope tractable. The two-alternative, forced-choice format is common in many applications of the DCE in outcomes research; many researchers, however, modify this format by including more than two alternatives, an opt-out alternative, or follow-up questions. Other researchers use other question formats (e.g., rating or ranking) or adaptive conjoint designs. Finally, other stated-preference techniques, such as best-worst scaling, have become increasingly popular in recent years. In many contexts, the methods illustrated here can be generalized and modified, but modeling other choice profiles or experiments can also involve complications. The task force believes that this report should be considered as a fundamental building block in making these issues (and their corresponding literature) more accessible for the average reader.

This report describes the results of alternative methods for analyzing data and provides some interpretation of the results of each type of analysis as applied to the pedagogical example. We also could have described extensions of the basic model, including interactions between attributes or incorporating respondent-specific information in the analysis. Furthermore, we did not fully address the ways in which these results could be used to calculate the relative importance of attributes, ratios capturing the trade-offs respondents are willing to make among changes in attribute levels (e.g., maximum acceptable risk or willingness to pay), or other measures of preference. Although each of these extensions of the basic model presented in this report is important, the task force agreed that an adequate treatment of these issues was beyond the scope of the present report.

Finally, like other task forces, we have drawn on a multitude of sources to compile this report, particularly the experiences of the task force members, the reviewer group, and the ISPOR members who have engaged in conversations and correspondence on this topic. The report is not meant to be a systematic review or collection of empirically verified results, but rather a description of present methods and a consideration of good research principles. We also acknowledge that the empirical and methodological literature on choice modeling continues to evolve, and new and refined analysis methods are expected to emerge. Therefore, this report may need to be updated or revised to reflect potentially changing norms and standards for the analysis of choice data.

Conclusions

This report describes commonly used methods for the analysis of DCE data and factors to consider when designing such an analysis. It presents researchers, reviewers, and readers of DCE studies with a basic understanding of the properties of the alternative methods and questions to consider when determining whether the analysis was appropriate for providing results that answer the research question the DCE was designed to assess. Although we have not aimed to identify best practices, we have intended to provide researchers with an understanding of the methods so that they can make informed decisions about the type of analysis method to use and to interpret the results accurately. Although the analysis methods presented in this report are those most commonly used to analyze DCE data in outcomes research, they are not the only methods that can be used, and readers are encouraged to keep abreast of emerging methods for analyzing these data.

Acknowledgments

The individual contribution by F. Reed Johnson is gratefully acknowledged. We thank the reviewers who commented during our forum at the ISPOR Montreal International Meeting and the ISPOR Milan European Congress. We especially thank the following who reviewed drafts of the report and submitted written comments: Kathy Beusterien, Marco Boeri, Brian Griner, Parul Gupta, Jinhai (Stephen) Huo, Christine Huttin, Jo Mauskopf, Brian Orme, Katherine Payne, Holly Peay, Shelby Reed, Dean Reynecke,

Lee Smolen, N. Udupa, Wendy Wan, and Leslie Wilson. Their feedback has both improved the article and made it an expert consensus ISPOR task force report. Finally, we thank Elizabeth Molsen for her assistance in developing this task force report, David Gebben for research support, Margaret Mathes for editorial support, and Lindsay Huey for graphics support.

Supplementary Materials

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jval.2016.04.004.

REFERENCES

[1] Bridges J, Hauber AB, Marshall D, et al. A checklist for conjoint analysis applications in health: report of the ISPOR Conjoint Analysis Good Research Practices Task Force. Value Health 2011;14:403–13.
[2] Johnson FR, Lancsar E, Marshall D, 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;16:3–13.
[3] Ryan M, Gerard K. Using discrete choice experiments to value health care programmes: current practice and future research reflections. Appl Health Econ Health Policy 2003;2:55–64.
[4] Bridges J, Kinter E, Kidane L, et al. Things are looking up since we started listening to patients: recent trends in the application of conjoint analysis in health 1970–2007. Patient 2008;1:273–82.
[5] Marshall D, Bridges J, Hauber AB, et al. Conjoint analysis applications in health—how are studies being designed and reported? An update on current practice in the published literature between 2005 and 2008. Patient 2010;3:249–56.
[6] de Bekker-Grob EW, Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. Health Econ 2012;21:145–72.
[7] Hauber AB, Fairchild AO, Johnson FR. Quantifying benefit-risk preferences for medical interventions: an overview of a growing empirical literature. Appl Health Econ Health Policy 2013;11:319–29.
[8] Clark MD, Determann D, Petrou S, et al. Discrete choice experiments in health economics: a review of the literature. Pharmacoeconomics 2014;32:883–902.
[9] Ryan M. Using conjoint analysis to take account of patient preferences and go beyond health outcomes: an application to in vitro fertilisation. Soc Sci Med 1999;48:535–46.
[10] Longworth L, Ratcliffe J, Boulton M. Investigating women's preferences for intrapartum care: home versus hospital births. Health Soc Care Community 2001;9:404–13.
[11] Hauber AB, Mohamed AF, Johnson FR, et al. Estimating importance weights for the IWQOL-Lite using conjoint analysis. Qual Life Res 2010;19:701–9.
[12] Mohamed AF, Hauber AB, Johnson FR, et al. Patient preferences and linear scoring rules for patient-reported outcomes. Patient 2010;3:217–27.
[13] Marshall D, McGregor SE, Currie G. Measuring preferences for colorectal cancer screening: what are the implications for moving forward? Patient 2010;3:79–89.
[14] Poulos C, Yang JC, Levin C, et al. Mothers' preferences and willingness to pay for HPV vaccines in Vinh Long Province, Vietnam. Soc Sci Med 2011;73:226–34.
[15] van Gils PF, Lambooij MS, Flanderijn MH, et al. Willingness to participate in a lifestyle intervention program of patients with type 2 diabetes mellitus: a conjoint analysis. Patient Prefer Adherence 2011;5:537–46.
[16] Bridges JF, Mohamed AF, Finnern HW, et al. Patients' preferences for treatment outcomes for advanced non-small cell lung cancer: a conjoint analysis. Lung Cancer 2012;77:224–31.
[17] Hauber AB, Arden NK, Mohamed AF, et al. A discrete-choice experiment of United Kingdom patients' willingness to risk adverse events for improved function and pain control in osteoarthritis. Osteoarthritis Cartilage 2013;21:289–97.
[18] Ho MP, Gonzalez JM, Lerner HP, et al. Incorporating patient-preference evidence into regulatory decision making. Surg Endosc 2015;29:2984–93.
[19] Sanders PMH, IJzerman MJ, Roach MJ, et al. Patient preferences for next generation neural prostheses to restore bladder function. Spinal Cord 2010;49:113–9.
[20] Groothuis-Oudshoorn CGM, Fermont JM, van Til JA, et al. Public stated preferences and predicted uptake for genome-based colorectal cancer screening. BMC Med Inform Decis Mak 2014;14:18.
[21] Plumb AA, Boone D, Fitzke H, et al. Detection of extracolonic pathologic findings with CT colonography: a discrete choice experiment of perceived benefits versus harms. Radiology 2014;273:144–52.
[22] Fishman J, O'Dwyer P, Lu HL, et al. Race, treatment preferences, and hospice enrollment: eligibility criteria may exclude patients with the greatest needs for care. Cancer 2009;115:689–97.
[23] Hall J, Kenny P, Hossain I, et al. Providing informal care in terminal illness: an analysis of preferences for support using a discrete choice experiment. Med Decis Making 2013;34:731–45.
[24] Nathan H, Bridges J, Cosgrove D, et al. Treating patients with colon cancer liver metastasis: a nationwide analysis of therapeutic decision-making. Ann Surg Oncol 2012;19:3668–76.
[25] Faggioli G, Scalone L, Mantovani LG, et al. Preferences of patients, their family caregivers and vascular surgeons in the choice of abdominal aortic aneurysms treatment options: the PREFER study. Eur J Vasc Endovasc Surg 2011;42:26–34.
[26] Arellano J, Hauber AB, Mohamed AF, et al. Physicians' preferences for bone metastases drug therapy in the United States. Value Health 2015;18:78–83.
[27] Morton RL, Snelling P, Webster AC, et al. Dialysis modality preference of patients with CKD and family caregivers: a discrete-choice study. Am J Kidney Dis 2012;60:102–11.
[28] Honda A, Ryan M, van Niekerk R, et al. Improving the public health sector in South Africa: eliciting public preferences using a discrete choice experiment. Health Policy Plan 2015;30:600–11.
[29] Regier DA, Peacock SJ, Pataky R, et al. Societal preferences for the return of incidental findings from clinical genomic sequencing: a discrete-choice experiment. CMAJ 2015;187:E190–7.
[30] Gray E, Eden M, Vass C, et al. Valuing preferences for the process and outcomes of clinical genetics services: a pilot study. Patient [Epub ahead of print] June 18, 2015.
[31] Johnson P, Bancroft T, Barron R, et al. Discrete choice experiment to estimate breast cancer patients' preferences and willingness to pay for prophylactic granulocyte colony-stimulating factors. Value Health 2014;17:380–9.
[32] Kauf TL, Yang JC, Kimball AB, et al. Psoriasis patients' willingness to accept side-effect risks for improved treatment efficacy. J Dermatolog Treat 2015;26:507–13.
[33] Whitty J, Filby A, Smith AB, et al. Consumer preferences for scanning modality to diagnose focal liver lesions. Int J Technol Assess Health Care 2015;31:27–35.
[34] Gonzalez JM, Johnson FR, Runken MC, et al. Evaluating migraineurs' preferences for migraine treatment outcomes using a choice experiment. Headache 2013;53:1635–50.
[35] Ryan M, Farrar S. Using conjoint analysis to elicit preferences for health care. BMJ 2000;320:1530–3.
[36] Bridges J. Stated-preference methods in health care evaluation: an emerging methodological paradigm in health economics. Appl Health Econ Health Policy 2003;2:213–24.
[37] Carlsson F, Martinsson P. Design techniques for stated preference methods in health economics. Health Econ 2003;12:281–94.
[38] Viney R, Savage E, Louviere J. Empirical investigation of experimental design properties of discrete choice experiments in health care. Health Econ 2005;14:349–62.
[39] Lancsar E, Louviere J. Conducting discrete choice experiments to inform healthcare decision making: a user's guide. Pharmacoeconomics 2008;26:661–77.
[40] Louviere J, Swait J, Hensher D. Stated Choice Methods: Analysis and Application. Cambridge, UK: Cambridge University Press, 2000.
[41] Hensher DA, Rose JM, Greene WH. Applied Choice Analysis: A Primer. Cambridge, UK: Cambridge University Press, 2005.
[42] Orme BK. Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research (2nd ed.). Madison, WI: Research Publishers LLC, 2010.
[43] McFadden D. Conditional logit analysis of qualitative choice behavior. In: Zarembka P, ed. Frontiers in Econometrics. New York, NY: Academic Press, 1974. p. 105–42.
[44] Cox DR, Snell EJ. Analysis of Binary Data (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC Press, 1989.
[45] Bech M, Gyrd-Hansen D. Effects coding in discrete choice experiments. Health Econ 2005;14:1079–83.
[46] Mark TL, Swait J. Using stated preference and revealed preference modeling to evaluate prescribing decisions. Health Econ 2004;13:563–73.
[47] Swait J, Louviere J. The role of the scale parameter in the estimation and comparison of multinomial logit models. J Market Res Soc 1993;30:305–14.
[48] McFadden D. Quantitative methods for analyzing travel behaviour on individuals: some recent developments. In: Hensher D, Stopher P, eds. Behavioral Travel Modelling. London: Croom Helm, 1978.

[49] Hess S, Rose JM. Can scale and coefficient heterogeneity be separated in random coefficients models? Transportation 2012;39:1225–39.
[50] Greene WH, Hensher DA. Does scale heterogeneity across individuals matter? A comparative assessment of logit models. Transportation 2010;37:413–28.
[51] Ben-Akiva M, Lerman SR. Discrete Choice Analysis: Theory and Application to Travel Demand. Cambridge, MA: MIT Press, 1985.
[52] Whitty JA, Ratcliffe J, Chen G, et al. Australian public preferences for the funding of new health technologies: a comparison of discrete choice and profile case best-worst scaling methods. Med Decis Making 2014;34:638–54.
[53] Allenby G, Rossi P. Marketing models of consumer heterogeneity. J Econom 1999;89:57–78.
[54] Allenby G, Lenk P. Modeling household purchase behavior with logistic normal regression. J Am Stat Assoc 1994;89:1218–31.
[55] Train K. Discrete Choice Methods with Simulation. New York, NY: Cambridge University Press, 2003.
[56] Huber J, Train KE. On the similarity of classical and Bayesian estimates of individual mean partworths. Mark Lett 2001;12:259–69.
[57] Regier DA, Ryan M, Phimister E, et al. Bayesian and classical estimation of mixed logit: an application to genetic testing. J Health Econ 2009;28:598–610.
[58] Dziak JJ, Coffman DL, Lanza ST, Li R. Sensitivity and specificity of information criteria. Methodology Center Technical Report #12-119. Penn State ScholarSphere. 2012. Available from: https://methodology.psu.edu/media/techreports/12-119.pdf. [Accessed March 19, 2016].