Issues in the Design of Discrete Choice Experiments

Introduction

The use of preference-elicitation tasks – in particular, discrete choice experiments (DCEs) – in Health Economics has grown significantly in recent decades (1). The most widely used DCE approach asks respondents to consider a series of hypothetical choices between alternatives (here called choice tasks), and to specify which alternative they prefer. The use of choice tasks in other areas – especially, psychology, transportation, marketing and agriculture – has a more established history. Health preference studies have been conduct for about as long (2, 3); however, the relatively late uptake in preference evidence in health is surprising in some regards as patient and population values concerning health have always been key components of a range of questions from health policy to clinical practice, and often cannot be directly observed, a problem exacerbated by the lack of a perfectly competitive market (4). Though there is broad consensus that the value patients or the population place on health matters in decision-making, the methods for including them in a way that is reliable are debated.

A significant issue with the conduct of such experiments is how best to construct the choice tasks to produce policy-relevant and reliable value estimates. If, for simplicity, the task has two alternatives (i.e., paired comparison (5)), which alternatives should be presented head-to-head? The risk of picking the wrong combinations are that values for some alternatives either cannot be estimated at all, or that they are estimated with an unacceptably low level of precision. This topic is of course not unique to health, and we should be cognisant of the work being conducted in other fields using similar methods. Conversely, we also believe that the design of health preference studies requires specific consideration to reflect the nature of the questions, and the provision of results, that best informs decision makers.

This paper provides a summary of a panel discussion from a DCE design symposium at the International Academy of Health Preference Research (IAHPR) 2018 meeting in Hobart, Australia on September 28, 2018. At the start of the symposium, each panellist presented lessons learned from their own experience:
John Rose (JR). A unified theory of experimental design for stated choice studies
Deborah J. Street (DJS). What can simulations tell us about DCE design performance?
Marcel Jonker (MJ). Individually adaptive D-efficient DCE designs
Paul Hansen (PH). The PAPRIKA method: A full factorial DCE involving pairwise rankings of all possible attribute combinations
Benjamin M. Craig (BMC). Experience-based methods for DCE designs

This summary has a dual role. It will lay out the key discussion points from this symposium, and will also identify important issues in design of DCEs passing along lessons to the broader community of health preference researchers. . After briefly defining design, we will talk about selecting respondents, about pre-identifying what is to be estimated, about tailoring the task to maximise data quality without over-burdening respondents, and about adapting DCE tasks to collect more preference evidence and estimate more precise values both at the individual and population level. Finally, we will conclude with some key areas of ongoing research which could be prioritised in coming years.

Design approaches

At the start of a health preference study, the researcher must first decide on the study's purpose and how to describe the alternatives, which is often based on qualitative research and contextual knowledge (6). The way we will discuss alternatives in a DCE is a combination of attributes and levels that are seen together by respondents. It is not feasible to assign every possible alternative and choice set to each respondent. DCE design refers to the selection and assignment of a manageable number of choice sets (e.g., which pairs of alternatives should be assigned to each respondent).

Selection of Choice Sets

Selection can be broken down into two sub-questions: which level combinations should be seen by respondents, and how should they be grouped into choice sets. Selection was the focus of the first two presentations at the symposium and an over-riding concern in DCE design has been described elsewhere (7). There are a variety of well-established methods for selecting choice sets, each making particular assumptions and with relative strengths and weaknesses.

In DJS's symposium presentation, a comparison was made between competing design approaches under two different prior assumptions about preferences. The construction approaches were: a generator-developed approach (8); a Modified Fedorov approach, which finds the best potential choice set to switch for each current choice set and makes the switch that most improves the design (9); and a coordinate exchange approach, which exchanges each level with every other levels, retaining those that improve the design (10). Importantly, most designs performed well in terms of both bias and accuracy.

JR's presentation argued that the distinction between design approaches is more about distinctions between underlying assumptions that each makes. After discussing the overlapping but not synonymous concepts of orthogonality, correlation and independence, JR presented the various optimality criteria that can be used to judge designs (discussed below). The key point from the presentation was that we as a field should focus on discussing the assumptions underlying the selection process rather than the merits of different design approaches per se.

Selection based on statistical efficiency

When selecting choice sets, it is important to note that assumptions about model type, likely preference structure, data coding, and the efficiency measure being optimised impact on the determination of the appropriate design. Either explicitly or implicitly, we have to pre-specify the coefficients that need to be estimable in the data analysis stage. Although most experiments will be interested in main effect terms, it is essential to consider how variables will be coded for analysis. A design based on the assumption that a variable will be treated as linear is unlikely to be optimal if that variable is dummy coded for analysis. Similarly, maybe experiments are interested in interactions, both between levels of different dimensions (does the level of X change the value of Y?), and between respondent demographics and levels of dimensions (do respondent groups X and Y value Z differently?). Also, it is important to pre-specify the regression technique to be used in analysis, as there are a number of competing options (see, for example (11)).

There is a variety of efficiency measures used in the DCE literature which are more or less suitable in different situations. D-error is commonly used in DCE design. This can be defined in terms of the determinant of the (asymptotic) variance-covariance matrix of the parameter estimates. Specifically, the design with the smallest possible determinant of the variance-covariance matrix is optimal, and efficiency of other designs is calculated relative to that lowest D-error.

The use of D-efficiency assumes that we are interested in the precision of regression coefficients, typically in a logit model. However, that is not necessarily the case and moving away from that assumption will impact on the relative value of different designs. Because we are usually interested in providing information for policy-makers in some natural units (such as a dollar-denominated willingness to pay), C-efficiency, focusing on some function of multiple parameters, often the ratio of coefficients reflecting willingness to pay, may be the preferred metric (12), although Kanninen noted that this approach is inherently inefficient (13). Similarly, depending on the circumstance, there is a range of other optimality criteria that might be considered, including the E-error (which minimises the variance of the least well-estimated contrast of unit length (see p.107 of Atkinson and Donev (14)), the G-error (minimises the variance of the predicted choice probabilities), or the S-error (minimises the necessary sample size). The important point is not that any of these measures are preferable to any other, simply that we have to be aware of how our data will most likely be used, and select choice sets based on the measures of efficiency that provides end-users with most certainty on the metrics they are interested in.

Assignment of Choice Sets

Though statistical efficiency is often very important in determining the suitability of a particular design, t issues concerning the assignment of choice sets are harder to measure, but also important.

Can the assignment of choice sets account for the preferences of a respondent or past respondents?

A frequently discussed issue is the appropriate way to integrate prior information about population preferences into the design process. This can take a variety of guises linked to the use of informative or non-informative priors. First, it is possible in the development of efficient designs to move away from specifying zero priors. The value of doing so is in maximising the precision of the regression coefficients in analysis. This particularly important when the study sample is small (either due to the population of interest being small or the sampling method being expensive) or when the research interest is pertaining to individual-level inference (rather than population-level inference).

It is sometimes the case that we do not have prior information about the extent (and in some situation even the direction) of the population preference. In that situation, it is often reasonable to soft-launch the survey under uninformative priors, estimate a model, then regenerate a design using the coefficients generated in this initial run of respondents. Unfortunately, the appropriate size of the soft launch is unknowable in advance, and will depend on both the likely distance of true preferences from the uninformative priors and the ultimate sample size available to the research team.

An extension of this concept was presented at the symposium by MJ. This work considered the use of individually adaptive D-efficient DCE designs in which the choice tasks for each individual respondent are optimized (in real-time during the survey administration) based on the respondents' observed choices. In several Monte Carlo simulation studies that were conducted, an increase in individual-level D-efficiency of approximately 25% to 40% was achieved compared to Bayesian efficient and near orthogonal DCE designs, respectively. Most importantly, only a few adaptive choice tasks at the end of the survey were sufficient to achieve most of these efficiency gains. The latter was also confirmed in a real-life DCE (which was hosted using a custom module integrated into Sawtooth Software).

Can the assignment of choice sets be tailored to the respondent's experience?

If attributes and levels are poorly understood or have different meanings to different respondents, focusing on statistical efficiency alone may lead to evidence of perception heterogeneity and uncertainty, not preferences. Experience-based designs ask respondents to choose between alternatives that they have experienced. Though the choices may be hypothetical, the alternatives are experienced, not described using vignettes.

BC's symposium presentation focused on the potential value of experience-based methods for DCE designs. In this work, the argument was advanced that decision-makers are likely to be more interested in the views of individuals with a clear understanding of what they are trading (e.g., patients). His presentation provided two examples, asking patients in a chemotherapy clinical to prioritize their symptom relief and asking patients who used two forms of delivery (infusion and infection) in a crossover trial about their preferences (15). The use of experience-based methods has a number of important advantages; in particular that it will reduce hypothetical bias. This has to be balanced against the potential for sample selection bias, small samples for rare alternatives, and a possible incompatibility between preferences of experienced individuals and the general population (16).

Can the assignment of choice sets be tailored to the respondents' capacity to respond accurately (i.e., respondent efficiency)?

Respondents often find choice tasks difficult, which can increase error variance and/or introduce the use of simplifying decision heuristics (17-19). This risk is likely to increase in a range of situations, particularly when the choice sets include many options, or have many dimensions and levels, or when each respondent is asked to complete a large number of choice tasks, or the respondent population is less cognitively capable of answering choice tasks, such as children, the elderly or people with cognitive difficulties. For a good empirical analysis of this issue, see Louviere *et al.* (20).

To attempt to mitigate this issue, it is important to consider ways to improve respondent efficiency. The usefulness of the approaches described below will often be context specific, but all have examples where they have been used successfully to improve ease of task completion. First, we can make the task easier by forcing a level of overlap on choice tasks. Experiments using partial profiles are in some ways an extreme version of this. More generally, we can set the design to have a fixed number of dimensions in each choice set that are the same in each of the options. Such a requirement can be imposed through a generator-type approach (by using generators with a fixed number of zeros), and also in the widely-used Ngene software (21). This is particularly valuable in situations where the number of dimensions is by necessity high (22). Albeit placing such constraints on the design generation process will, in isolation, worsen the design's statistical efficiency (holding the measure used constant), the question is whether this is justified by an improvement in respondent experience (and hopefully comprehension and engagement). This trade-off has to be considered in the context of the expected sample size in the survey. The larger the sample size, the more consideration can be given to respondent efficiency. If constraints are placed on a design which reduces efficiency under some set of priors, and that design is then used in a large enough sample, we as analysts can still produce point estimates that are precise enough for any practical use.

The researcher will generally be able to pre-specify the number of choice tasks that each individual will complete. Again, this decision reflects a balance between two competing concerns, namely maximising the quantity of data, and respondent ease of completion. The Clark review identifies the range of questions asked of each respondent in health-based choice tasks (1). What the optimal number of tasks is, is again likely to be partially context specific, so it is worthwhile to consider a number of ways of assessing whether later questions provide more information or simply more frustration and noise. A recent review grouped these approaches into four main categories, namely tests of measurement validity, measurement reliability, choice validity, and choice reliability (23).

PH's symposium presentation concerned the ***P****otentially* **A***ll* **P***airwise* **R***an***K***ings of all possible* **A***lternatives* (PAPRIKA) method (24), and described its implementation in the 1000minds software. This approach differs from conventional DCEs, albeit it is widely used for DCE applications, especially in health and Multi-Criteria Decision-Making (25). Central to the PAPRIKA method is the fundamental principle that an *overall* ranking of all profiles representable by a given set of atttributes and levels – i.e. all possible combinations of the levels on the attributes – is defined when all *pairwise* rankings of the partial profiles vis-à-vis each other are known (and provided the rankings are consistent).

The PAPRIKA method's closest theoretical antecedent is Pairwise Trade-off Analysis (PTA) developed by Rich Johnson in the 1970s (16), and a precursor to Adaptive Conjoint Analysis (26). Like PTA, PAPRIKA is based on the idea that undominated pairs – pairs of partial profiles where one has a higher ranked level for at least one attribute and a lower ranked level for at least one other attribute than the other profile – that are explicitly ranked by the decision-maker can be used to implicitly rank other undominated partial pairs. In most applications, the PAPRIKA method simply asks respondents to pairwise rank partial profiles differing with respect to only two attributes at a time, which is advantageous in terms of respondent ease. Through use of dominance and transitivity, the approach infers potentially all pairwise rankings of all possible profiles (24). From these pairwise rankings, part-worth utilities are obtained via linear programming.

Conclusions

The increasing relevance of health preference research has led to methodological innovations and discoveries regarding the selection and assignment of choice set. Interested researchers may consider using simulation techniques prior to their study or to explore their assumptions after data collection. They represent a low-cost method for getting information about how designs will perform in the field, identifying errors before the research team commits financial and time resources to broad data collection. In the absence of simulation studies preceding DCE fieldwork, we recommend that researchers consider sample size requirements to generate adequately precise parameter estimates (27). After their study, simulation may show differences between choice-set blocks, sample sizes and other design assumptions.

If a researcher is fortunate to have a large sample size available and predominantly interested in population level inference, choice set selection is mostly about identification. Therefore, DCE designs that do not make use of informative priors are likely to yield adequately precise estimates, as long as they are identified and have sufficient variation. Patient preference studies often have smaller sample sizes which motivates the use informative priors, adaptive approaches or simplified tasks (e.g., experience-based or partial profiles). In case of extremely small samples, it may be worthwhile to consider individually adaptive, efficient choice tasks, which show considerably early promise, but are more assumption laden. However, these techniques are currently not well-established.

Separate from statistical efficiency is the aspect of respondent (or behavioural) efficiency. That is, statistical efficiency alone should not be the only design goal; in many instances, it is reasonable and indeed preferable to deliberately reduce the level task complexity at the expense of statistical efficiency, such as the clinical trial described by BMC. If a survey is to be administered in a cognitively challenged population, then consideration should be made of the potential use of overlapping designs and any presentational approaches to make the respondents' task more straightforward.

1. Clark MD, Determann D, Petrou S, Moro D, de Bekker-Grob EW. Discrete choice experiments in health economics: a review of the literature. Pharmacoeconomics. 2014;32(9):883-902.
2. Thorndike EL. Valuations of certain pains, deprivations, and frustrations. The Pedagogical Seminary and Journal of Genetic Psychology. 1937;51(2):227-39.
3. Thurstone LL. The Method of Paired Comparisons for Social Values. Journal of Abnormal and Social Psychology. 1927;21:384-400.
4. Arrow KJ. Uncertainty and the welfare economics of medical care. Am Econ Rev. 1963;53(5):941-73.
5. David HA. The Method of Paired Comparisons. New York: Hafner Publishing Company; 1963. 124 p.
6. Coast J, Al-Janabi H, Sutton EJ, Horrocks SA, Vosper AJ, Swancutt DR, et al. Using qualitative methods for attribute development for discrete choice experiments: issues and recommendations. (doi: 10.1002/hec.1739.). Health Econ. 2012.
7. Reed Johnson F, Lancsar E, Marshall D, Kilambi V, Muhlbacher A, Regier DA, et al. Constructing experimental designs for discrete-choice experiments: report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force. Value Health. 2013;16(1):3-13.
8. Street DJ, Burgess L. The Construction of Optimal Stated Choice Experiments: Theory and Methods. Hoboken, New Jersey: Wiley; 2007.
9. Cook RD, Nachtsheim CJ. A comparison of algorithms for constructing exact D-optimal designs. Technometrics. 1980;22(3):315-24.
10. Meyer RK, Nachtsheim CJ. The Coordinate-Exchange Algorithm for Constructing Exact Optimal Experimental Designs. Technometrics. 1995;37(1):60-9.
11. Fiebig D, Keane M, Louviere J, Wasi N. The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity. Marketing Science. 2010;29(3):393-421.
12. Scarpa R, Rose JM. Design efficiency for non-market valuation with choice modelling: how to measure it, what to report and why. Australian Journal of Agricultural and Resource Economics. 2008;52(3):253-82.

13. Kanninen B, editor Optimal Design of Choice Experiments for Non-Market Valuation. Stated Preference: What do we know? Where do we go?; 2000; Washington DC.
14. Atkinson AC, Donev AN. Optimum Experimental Designs: Oxford Science Publications; 1992.
15. Rummel M, Kim TM, Aversa F, Brugger W, Capochiani E, Plenteda C, et al. Preference for subcutaneous or intravenous administration of rituximab among patients with untreated CD20+ diffuse large B-cell lymphoma or follicular lymphoma: results from a prospective, randomized, open-label, crossover study (PrefMab). Ann Oncol. 2017;28(4):836-42.
16. Brazier J, Ratcliffe J, Salomon JA, Tsuchiya A. Measuring and valuing health benefits for economic evaluation. Oxford: Oxford University Press; 2007.
17. Erdem S, Campbell D, Hole AR. Accounting for attribute-level non-attendance in a health choice experiment: Does it matter? . Health Econ. 2014.
18. Hole AR, Norman R, Viney R. Response Patterns in Health State Valuation Using Endogenous Attribute Attendance and Latent Class Analysis. Health Econ. 2014.
19. Jonker MF, Donkers B, De Bekker-Grob EW, Stolk EA. The Effect of Level Overlap and Color Coding on Attribute Non-attendance in Discrete Choice Experiments. Value Health. 2017.
20. Louviere JJ, Islam T, Wasi N, Street D, Burgess L. Designing Discrete Choice Experiments: Do Optimal Designs Come at a Price? Journal of Consumer Research. 2008;35(2):360-75.
21. Choice Metrics Pty Ltd. Ngene User Manual and Reference Guide (version 1.2). 2018.
22. Norman R, Viney R, Aaronson NK, Brazier JE, Cella DF, Costa DSJ, et al. Using a discrete choice experiment to value the QLU-C10D: Feasibility and sensitivity to presentation format. Qual Life Res. 2016.
23. Janssen EM, Marshall DA, Hauber AB, Bridges JFP. Improving the quality of discrete-choice experiments in health: how can we assess validity and reliability? Expert Review of Pharmacoeconomics and Outcomes Research. 2017;17(6):531-42.
24. Hansen P, Ombler F. A new method for scoring multi-attribute value models using pairwise rankings of alternatives. Journal of Multi-Criteria Decision Analysis. 2008;15:87-107.
25. Thokala P, Devlin N, Marsh K, Baltussen R, Boysen M, Kalo Z, et al. Multiple Criteria Decision Analysis for Health Care Decision Making--An Introduction: Report 1 of the ISPOR MCDA Emerging Good Practices Task Force. Value Health. 2016;19(1):1-13.
26. Green PE, Krieger AM, Wind Y. Thirty years of conjoint analysis: Reflections and Prospects. Interfaces. 2001;31(3):S56.
27. de Bekker-Grob EW, Donkers B, Jonker MF, Stolk EA. Sample Size Requirements for Discrete-Choice Experiments in Healthcare: a Practical Guide. Patient. 2015;8(5):373-84.