# OPTIMAL DESIGNS AND ANALYSES FOR DISCRETE CHOICE EXPERIMENTS

# DETERMINING OPTIMAL DESIGNS AND ANALYSES FOR DISCRETE CHOICE EXPERIMENTS

**By:**

**THUVARAHA (THUVA) VANNIYASINGAM, BSc, MSc**

**A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Doctor of Philosophy**

**McMaster University DOCTOR OF PHILOSOPHY (2014) Hamilton, Ontario**

**(Health Research Methodology – Clinical Epidemiology Specialization)**

**Title: Design and analysis of discrete choice experiments to optimize decision-making**

**AUTHOR: Thuvaraha (Thuva) Vanniyasingam, BSc (McMaster University, 2011),**

**MSc (McMaster University, 2013)**

**SUPERVISOR: Professor Lehana Thabane**

**NUMBER OF PAGES: xiv; 120**

## LAY ABSTRACT

This thesis focuses on the design and analysis of preference surveys, which are referred to as discrete choice experiments. These surveys are used to capture and quantify individuals' preferences on various characteristics describing a product or service. They are applied in various health settings to better understand a population. For example, clinicians may want to further understand a patient population's preferences in regards to multiple treatment alternatives. Currently, there is no optimal approach for designing or analyzing preference surveys. We investigated what factors help improve the design of a preference survey by exploring the literature and conducting our own simulation study. We also investigated how sensitive the results of a preference survey were based on the statistical model used. Overall, we found that (i) increasing the amount of information presented and reducing the number of variables to explore will maximize the statistical optimality of the survey; and (ii) analyzing the data with different statistical models will yield similar results in the ranking of individuals' preferences of the variables explored.

# ABSTRACT

**Background and Objectives:**

Understanding patient and public values and preferences is essential to healthcare and policy decision making. Discrete choice experiments (DCEs) are a common tool used to capture and quantify these preferences. Recent technological advances allow for a variety of approaches to create and analyze DCEs. However, there is no optimal DCE design, nor analysis method.

Our objectives were to (i) survey DCE simulation studies to determine what design features affect statistical efficiency, and assess their reporting, (ii) further investigate these findings with a *de novo* simulation study, and (iii) explore the sensitivity of individuals' preference of attributes to several methods of analysis.

**Methods:**

We conducted a systematic survey of simulation studies within the health literature, created a DCE simulation study of 3204 designs, and performed two empirical comparison studies. In one empirical comparison study, we determined addiction agency employees' preferences on knowledge translation attributes using four models, and in the second, we determined elementary school children's choice of bullying prevention programs using nine models.

**Results and Conclusions:**

In our evaluation of DCE designs, we identified six design features that impact the statistical efficiency of a DCE, several of which were further investigated in our simulation study. The reporting quality of these studies requires improvement to ensure that appropriate inferences can be made, and that they are reproducible.

In our empirical comparison of statistical models to explore the sensitivity of individuals preferences of attributes, we found similar rankings in the relative importance measures of attributes' mean part-worth utility estimates, which differed when using latent class models.

Understanding the impact of design features on statistical efficiency are useful for designing optimal DCEs. Incorporating heterogeneity in the analysis of DCEs may be important to make appropriate inferences about individuals' preferences of attributes within a population.

## ACKNOWLEDGEMENTS

This thesis would not have been possible without the support and guidance of the many incredible individuals whom I am grateful to have in my life.

I would first like to thank Dr. Lehana Thabane for his mentorship, guidance, and support throughout this entire process. You were the first person in my life to see my true potential and push me to see beyond my own glass ceilings. Over the years, you have consistently gone above and beyond as a PhD supervisor. You are an outstanding researcher with a high standard of integrity, honesty, and incredible work ethic. As a mentor, you have been a friend during the lows, a cheerleader during the highs, and a strong guide when I strayed off course. I have truly flourished through your support and this entire process. Thank you for always being available to answer my questions and concerns, and for helping me gain the confidence I need to now pursue anything in my life. My experience under your mentorship is invaluable and I truly cannot thank you enough for the massive impact you have made on my life.

I would like to thank my thesis committee, Drs. Gary Foster and Charles E. Cunningham. You have both always been so thoughtful, patient, and ready to provide advice, assistance, and resources whenever I needed it. Thank you for always being so encouraging and willing to share your time, knowledge, and wisdom.

To Dr. Lawrence Mbuagbaw, you have been a cornerstone to me during this entire experience. Your mentorship has been critical to my success in this PhD, and to my development as an independent researcher. Thank you for recognizing my challenges and taking time out of your busy schedule just to sit down and teach me different aspects of research, without me asking or even sometimes realizing I needed help. Thank you for checking in on me to make sure I was

doing okay when you didn't hear from me for a few weeks. I have so much gratitude for your endless generosity with time, your key motivational phrases (that are now engraved in my memory), and - what I cherish most - your friendship.

I would like to thank my colleagues and the staff in the Department of Health Research Methodology, Evidence, and Impact who helped make my journey feel supported at all levels, collaborative, and encouraging.

To my parents, brother, sister, and beautifully massive, extended family: you have been so considerate and patient with me. Thank you for being there for me as I spent so much of my time pursuing my ambitions. Your love and support in my life is invaluable, and I cannot wait to spend more time with you in the next chapter of my life.

Lastly, to my best friends who have been there with me through everything, I cannot thank you enough for the support. From the phone calls and the nights out, to the conversations about self-doubt, self-care, and self-motivation, you have helped carry my spirit through this journey. You have been, and continue to be, my rock.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

## LIST OF ALL ABBREVIATIONS

| | |
|---|---|
| BIC | Bayesian information criterion |
| CL | Conditional logit |
| DCE | Discrete choice experiment |
| D-efficiency | Design efficiency |
| D-error | Design error |
| DIC | Deviance information criterion |
| D-optimality | Design optimality |
| ECL | Error components logit |
| EQ-5D | EuroQol five dimensions |
| ESS | Estimated sample size |
| GCLASS | Group-level latent classes |
| HPD | Highest posterior density |
| HUI | Health utility index |
| ICC | Intraclass correlation coefficient |
| IIA | Independence of irrelevant alternatives |
| LL | Log-likelihood |
| LR | Literature review |
| MNL | Multinomial logit |
| MNP | Multinomial probit |
| RE | Random effects |
| RI | Relative importance |
| RPL | Random parameter logit |
| SAS | Statistical analysis software |
| SR | Systematic review |

## DECLARATION OF ACADEMIC ACHIEVEMENT

This thesis is a sandwich thesis which combines four individual projects prepared for publication in peer-reviewed journals. Two have been published and one has been so. Thuva Vanniyasingam has contributed to the following components in all of the studies: developing the research questions, writing the protocols and statistical analyses plans, data extraction (for two of the four studies) and management, conducting the statistical analyses; designing the figures; writing all manuscripts; submitting the manuscripts and responding to reviewers' comments. The work in this thesis was conducted between Fall 2012 and Summer 2018.

# CHAPTER 1

# INTRODUCTION

## 1.0 Capturing preferences for clinical care

Health care increasingly includes multiple stakeholders (clinicians, patients, caregivers, policy makers, communities) in making decisions about individual patient care and health care services[1]. By eliciting and integrating their preferences into the health care, we optimize clinical decision making and the development of clinical practice guidelines that include patient-optimal outcomes [2]. Individuals' preferences are derived from their evaluation of various dimensions of a service, therapy, or health outcome[2]. Preferences are based on an individual's perspectives, priorities, beliefs, expectations, values and goals [3].

With recent technological advancements, several computer-based applications have surfaced to capture stakeholder preferences. In an overview of 22 systematic reviews of values and preferences for guideline development, a variety of elicitation methods were described[4]. Zhang and colleagues categorize these methods into three segments: a) qualitative information, b) utility or health status value, and c) non-utility, quantitative information[4]. The qualitative information category includes focus groups, interviews, and participant or non-participant observation[5]. The utility or health status value category includes techniques that quantify preferences using standard gamble[6], time trade off[7-10], probability trade-off techniques [11-14], visual analogue scale[15-17], and multi-attribute instruments such as the EuroQol five dimensions questionnaire (EQ-5D utility)[18] and health utility index (HUI utility)[19]. The non-utility, quantitative information category includes methods to elicit participants' preferences by either asking them to choose from a set of options or by presenting a non-utility measurement of health states. Examples of these include discrete choice experiments [20, 21] and self-developed questionnaires and scales[22, 23].

This dissertation is focused on discrete choice experiments (DCEs), a quantitative technique for eliciting preferences.

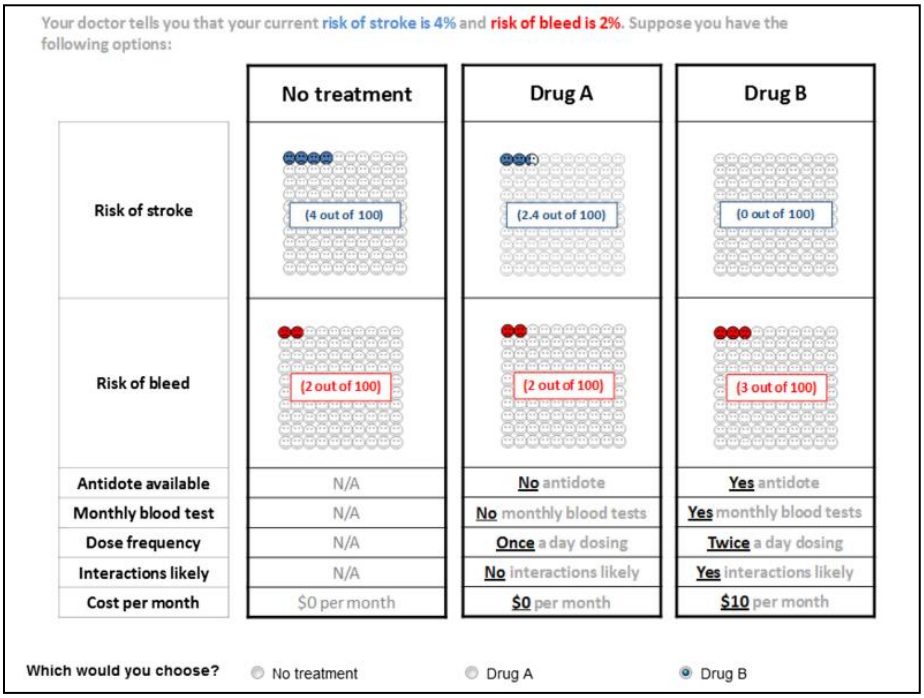
## 2.0 An introduction to discrete choice experiments (DCEs)

Since the 1990s, there has been a rapid increase in the use of DCEs to measure the preferences of stakeholders, most commonly healthcare workers, patients, and members of the public. Applications include health services research to investigate human resource issues related to health workers (nurses, doctors, pharmacists, policy makers, and clinical officers)[24-30], understanding patient preferences to inform therapies or treatments [31-35], and incorporating public preferences on health services [36-38]. Several systematic and literature reviews have summarized the evidence of the vast quantity of DCEs in each area of health research [29, 39-44].

In a DCE, participants are asked to choose between alternative health care options that differ in characteristics (or attributes) [45]. The option can be a health product (such as a treatment or therapy approach), program, policy, or service. Investigators must determine the attributes that best characterize this component [46], and each attribute's range of values (or levels), using focus groups or surveys. For example, for the component of oral anticoagulant medication, participants choose between drug options that are characterized by seven attributes, including risk of stroke, antidote, and dose frequency [47]. Levels of attributes are varied within alternative components. In the same example, attribute levels for risk of stroke included 0%, 0.8%, 1.6%, and 2.4%, antidote included yes or no, and dose frequency included once or twice per day[47]. Each question (choice task) presents two or more alternatives - which vary by these attribute levels - for participants to choose from. One choice task's alternative can be a drug with 2.4% risk of stroke, no antidote, and once per day dose frequency. This can be presented with a no drug alternative or

a drug alternative with 0% risk of stroke, an antidote, and a twice per day dose frequency. Figure 1 presents an illustration of this choice task example and was obtained from Ghibjen and colleagues' study [47]:

Figure 1: Example of a choice task




Based on the alternatives presented in a choice task, participants choose their most preferred drug alternative, and this process is repeated for several choice tasks in a survey.

When designing a DCE, investigators must consider the number of choice tasks, alternatives, attributes, and levels within attributes. A DCE must have a minimum of two alternatives per choice task. Beyond this, there is no limit to the number of alternatives and the number of choice tasks one can include in a DCE. A full factorial design will present all possible combinations of attributes and attribute levels, however this is not always possible. For studies with a large number of attributes and attribute levels, as many as billions of possible combinations can be created. For example, a DCE with 16 four-level attributes has over 4 billion different types

of alternatives, each with a unique combination of attributes ($4^{16}$=4,294,967,296). When there is a high volume of possible combinations, it is only feasible to present a fraction of these alternatives in a survey, termed a fractional factorial design.

When analyzing a DCE, investigators use participants' choice task responses to estimate the strength of their preferences regarding changes in attribute levels [45]. They are commonly referred to as "part-worth utilities," or "preference weights," and are estimated using regression models[45]. These part-worth utility estimates can also be transformed into relative importance measures, to describe the trade-offs of each attribute based on the range of levels included in the study [45]. When a fractional factorial design is the only feasible option, statistical considerations are important when making inferences about participants' choices.

### 2.1 Statistical considerations for DCEs

Several methodological considerations are required to ensure the part-worth utility estimates generate relative importance measures that are representative of participants' preferences. A poorly designed DCE may lead to poor data quality, resulting in low reliability of statistical estimates or erroneous conclusions[48]. A DCE must include both an optimal design and statistical analysis approach, in order to maximize the precision of parameter estimates for a given number of choice tasks[49].

At the design stage, there are two components to an efficient design - statistical efficiency and informant (or response) efficiency. Statistical efficiency is a function of how equally each level of an attribute, and each pair of levels, appear within the design[46]. Maximizing statistical efficiency will minimize the width of confidence intervals around parameter estimates - in this case part-worth utility estimates - for a given sample size[46]. Response efficiency is a

measurement error, and is influenced by participants' inattention to the choice tasks or other unobserved influences[46]. A couple of examples is when participants are burdened with answering a large number of choice tasks or from choosing from a large number of alternatives per choice task. To maximize the precision of estimates, investigators must consider study-design trade-offs to optimize both statistical efficiency and response efficiency[46]. A full-factorial design presents all possible alternatives and thus has maximum statistical efficiency, however, this is at the expense of response efficiency. Fractional factorial designs are more attractive, because they reduce participant burden, and thus increase response efficiency. With only a portion of all possible alternatives presented, however, it is critical to assess the statistical efficiency of fractional factorial designs.

At the statistical analysis stage, determining an appropriate model is critical for accuracy and inference. A variety of statistical approaches are used to model DCE data, including weighted least squares method, fixed effects logit or probit models, random effects or mixed effects logit or probit models, hierarchical Bayesian models, and latent class models [45, 50]. It is important to consider potential correlations within the data, because participants' responses to several choice tasks are likely to be similar. Models that do not appropriately capture within- and between-participant variability are limited in their ability to adequately characterize heterogeneity in mean preferences[51]. Ignoring these clustering effects in the analysis stage can potentially bias the part-worth utility estimates and relative importance measures of attributes, resulting in misleading inferences.

## 2.2 Strengths of DCEs

Discrete choice experiments (DCEs) are a relatively easy and inexpensive approach to determining the relative importance of attributes involved in making complex decisions related to health outcomes and health care services [52-67]. They are commonly used in health research for eliciting participants' preferences on components with multiple attributes and levels within attributes, to provide quantitative estimates of preferences. DCEs allow us to explore multiple factors in addition to health outcomes, such as waiting time, location of treatment, type of care, and staff providing care[68]. Investigators can estimate participants' trade-offs between these process-type attributes and health outcome attributes, to understand more about what individuals prioritize when making health care decisions[68]. DCEs have high levels of internal validity and convergent validity compared to standard gamble and willingness-to-pay approaches [55, 69].

## 3.0 Methodological challenges of DCEs addressed in this thesis

The general objectives of this thesis are to (i) summarize the literature about the impact of design features on statistical efficiency and assess their reporting quality, (ii) use simulations to determine optimal designs that enhance statistical efficiency, and (iii) explore the sensitivity of attribute rankings to the method of analysis. These objectives are intended to address three key challenges at the design, analysis, and reporting stages of DCEs.

### 3.1 Summarizing the evidence and appraising the reporting quality of DCE simulation studies

Understanding the impact of DCE design characteristics on statistical efficiency will bring more power to investigators during the design stage. They can reduce the variance of estimates by manipulating their designs to construct a simpler DCE that is statistically efficient and minimizes participants' response burden. Currently there are several studies exploring DCE designs. These studies range from comparing or introducing new statistical optimality criteria[70, 71] to approaches for generating DCEs[72], to exploring the impact of different prior specifications on parameter estimates[73-75]. To our knowledge, the results of these findings have not been summarized. This may be due to a variation in objectives and outcomes across studies or poor reporting quality, making it difficult to synthesize information and draw conclusions. Poor reporting quality also makes it difficult to reproduce the simulations.

In this thesis, we systematically surveyed simulation studies in the health literature to determine the impact of design features on statistical efficiency—measured using relative D-efficiency, relative D-optimality, or D-error. We also appraised the completeness of reporting of the studies using the criteria for reporting simulation studies[76].

### 3.2 Designing an optimal DCE

There is no optimal DCE design. Investigators must balance both statistical efficiency and response efficiency. For DCE designs exploring a large number of variables, where presenting all combinations of alternatives is not feasible, a fractional factorial design can be used to determine individuals' preferences. When a small fraction of all possible scenarios is used in a DCE, biased results may occur if attributes and attribute levels are not evenly represented. Previous studies have taken various directions to explore statistical efficiency, either empirically or with simulated data. These approaches (i) identified optimal designs using specific design characteristics [77-79], (ii) compared different statistical optimality criteria [71, 80], (iii) explored prior estimates for Bayesian designs [75, 81-83], and (iv) compared designs with different methods to construct choice tasks (such as random allocation, swapping, cycling, etc.) [80, 84-87]. Detailed reports have been produced to describe the key concepts behind DCEs such as their development, design components, statistical efficiency, and analysis [49, 88]. However, these reports did not address the effect of having more attributes or more alternatives on efficiency.

To our knowledge, no study has investigated the impact of multiple DCE characteristics with pragmatic ranges on statistical efficiency. In this thesis, we conducted a simulation study to investigate how statistical efficiency, measured with relative D-efficiency, D-optimality, or D-error, is influenced by various experimental DCE design characteristics including the number of: choice tasks, alternatives, attributes, and attribute levels.

### 3.3 Impact of clustering on DCEs analyses

While DCEs are able to quantify individuals' preferences, correlated responses may arise when each participant is asked to answer more than one choice task. For feasibility, each participant often completes an entire survey, making choices on several choice tasks. As more than one choice task are completed by each individual, their responses are likely to be similar, potentially biasing the utility measures. Models that do not appropriately capture within- and between- participant variability are limited in their ability to adequately characterize heterogeneity in mean preferences[51]. Participants may also belong to observed or unobserved (latent) groups. Traditional fixed effects models, such as a conditional logit or multinomial logit models, assume there are no correlations within the data. Several studies used these models for analyzing DCE data, which may not have been appropriate due to potential clustering effects [58, 89-92]. This creates a challenge in the interpretation of the results of these studies, as we are unsure of how sensitive the results are to the hierarchical nature of DCE data. Understanding the impact of clustering effects on relative importance measures and the ranking of attributes is critical for ensuring that the interpretation of DCE results is accurate, particularly since a variety of adjusted and unadjusted models are currently being used.

In this thesis, we conducted two empirical comparisons to determine how robust the rankings of attributes' relative importance measures were across various hierarchical model settings, using two empirical survey datasets.

**4.0 Scope of the thesis**

This thesis is a "sandwich" of four papers. First, we reviewed simulation studies of DCEs to determine how survey design features affect statistical efficiency. Second, we investigated the impact of various DCE designs on statistical efficiency in a simulation study. Third, we conducted an empirical comparison study to determine whether the final rankings of attributes were sensitive to clustering effects in using 1- and 2-level models. Fourth, we conducted a similar empirical comparison study, this time also using 3-level models.

We address the following research questions:

1. What do DCE simulation studies conclude on the impact of various design features on relative design efficiency?
2. What is the quality of reporting of DCE simulation studies?
3. How will varying the number of choice tasks, alternatives, attributes, and levels within attributes impact the statistical efficiency of a DCE?
4. How robust is the ranking of attributes when different fixed effects and random effects approaches are used to analyse DCE data — namely, multinomial logit (MNL) and multinomial probit (MNP) models?
5. How robust is the ranking of attributes when various approaches are used to adjust for multi-level clustering — namely, fixed effects conditional logit model, latent class models, and hierarchical latent class models?

Investigating these questions will lead to a compelling body of evidence that will inform clinicians, researchers, decision makers and researchers on how to best design and analyze DCEs.

Chapter 2 is divided into two parts. First, we systematically surveyed DCE simulation studies to determine how survey design features affect statistical efficiency. Statistical efficiency was

measured using relative design (D-) efficiency, D-optimality, or D-error. Second, we appraised study reporting quality using the criteria for reporting simulation studies.

Chapter 3 further explores several key design features identified in the systematic survey in a *de novo* simulation study on relative D-efficiency. A variety of fractional factorial designs were created to identify optimal approaches for creating DCEs.

Chapters 4 and 5 focus on empirical comparisons of statistical methods used to analyze DCE data. The relative importance measures of attributes were determined from the part-worth utility estimates derived from various regression models. These relative importance measures were then ranked to inform investigators which attributes were most preferred. A variety of approaches to adjust for potential clustering effects within DCE data were explored.

In Chapter 4, service providers and administrators of addiction agencies for women were surveyed to determine what knowledge dissemination attributes were of most value to them. The goal of this study was to investigate the impact of the clustering of participant responses by assessing the robustness of the ranking of attributes. Four one-level and two-level models were explored: a fixed-effects multinomial logit (MNL) model, fixed effects multinomial probit (MNP) model, a random effects MNL model, and a random effects MNP model.

In Chapter 5, elementary school students were asked for their choice of bullying prevention programs. This data had five potential layers of clustering effects: choice tasks, individuals, classrooms, grades, and schools. We explored a fixed conditional logit model and several latent class and hierarchical latent class models. The clusters were adjusted as latent classes, covariates, or random effects. [93]

**References**

1. Dirksen CD: **The use of research evidence on patient preferences in health care decision-making: issues, controversies and moving forward**. *Expert review of pharmacoeconomics & outcomes research* 2014, **14**(6):785-794.
2. Brennan PF, Strombom I: **Improving health care by understanding patient preferences: the role of computer technology**. *Journal of the American Medical Informatics Association* 1998, **5**(3):257-262.
3. Montori VM, Elwyn G, Devereaux PJ, Straus SE, Haynes RB, Guyatt G: **Decision Making and the Patient**. In: *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice, 3rd ed.* edn. Edited by Guyatt G, Rennie D, Meade MO, Cook DJ. New York, NY: McGraw-Hill Education; 2015.
4. Zhang Y, Coello PA, Brożek J, Wiercioch W, Etxeandia-Ikobaltzeta I, Akl EA, Meerpohl JJ, Alhazzani W, Carrasco-Labra A, Morgan RL: **Using patient values and preferences to inform the importance of health outcomes in practice guideline development following the GRADE approach**. *Health and quality of life outcomes* 2017, **15**(1):52.
5. Gooberman-Hill R: **Qualitative approaches to understanding patient preferences**. *The Patient: Patient-Centered Outcomes Research* 2012, **5**(4):215-223.
6. Gafni A: **The standard gamble method: what is being measured and how it is interpreted**. *Health services research* 1994, **29**(2):207-224.
7. Guo J, Konetzka RT, Dale W: **Using time trade-off methods to assess preferences over health care delivery options: a feasibility study**. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2014, **17**(2):302-305.

8. Gu NY, Wolf C, Leopold S, Manner PA, Doctor JN: **A comparison of physician and patient time trade-offs for postoperative hip outcomes**. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2009, **12**(4):618-620.

9. Protheroe J, Fahey T, Montgomery AA, Peters TJ: **The impact of patients' preferences on the treatment of atrial fibrillation: observational study of patient based decision analysis**. *BMJ (Clinical research ed)* 2000, **320**(7246):1380-1384.

10. Dranitsaris G, Stumpo C, Smith R, Bartle W: **Extended dalteparin prophylaxis for venous thromboembolic events: cost-utility analysis in patients undergoing major orthopedic surgery**. *American journal of cardiovascular drugs : drugs, devices, and other interventions* 2009, **9**(1):45-58.

11. Alonso-Coello P, Montori VM, Sola I, Schunemann HJ, Devereaux P, Charles C, Roura M, Diaz MG, Souto JC, Alonso R *et al*: **Values and preferences in oral anticoagulation in patients with atrial fibrillation, physicians' and patients' perspectives: protocol for a two-phase study**. *BMC Health Serv Res* 2008, **8**:221.

12. Devereaux PJ, Anderson DR, Gardner MJ, Putnam W, Flowerdew GJ, Brownell BF, Nagpal S, Cox JL: **Differences between perspectives of physicians and patients on anticoagulation in patients with atrial fibrillation: observational study**. *BMJ (Clinical research ed)* 2001, **323**(7323):1218-1222.

13. Man-Son-Hing M, Laupacis A, O'Connor AM, Coyle D, Berquist R, McAlister F: **Patient preference-based treatment thresholds and recommendations: a comparison of decision-analytic modeling with the probability-tradeoff technique**. *Medical*

*decision making : an international journal of the Society for Medical Decision Making* 2000, **20**(4):394-403.

14. Man-Son-Hing M, Hart RG, Berquist R, O'Connor AM, Laupacis A: **Differences in treatment preferences between persons who enrol and do not enrol in a clinical trial**. *Annals (Royal College of Physicians and Surgeons of Canada)* 2001, **34**(5):292-296.

15. Fried TR, Tinetti M, Agostini J, Iannone L, Towle V: **Health outcome prioritization to elicit preferences of older persons with multiple health conditions**. *Patient Educ Couns* 2011, **83**(2):278-282.

16. Wang Y, Xie F, Kong MC, Lee LH, Ng HJ, Ko Y: **Patient-reported health preferences of anticoagulant-related outcomes**. *J Thromb Thrombolysis* 2015, **40**(3):268-273.

17. Sun C, Brown AJ, Jhingran A, Frumovitz M, Ramondetta L, Bodurka DC: **Patient preferences for side effects associated with cervical cancer treatment**. *International journal of gynecological cancer : official journal of the International Gynecological Cancer Society* 2014, **24**(6):1077-1084.

18. Goudarzi R, Zeraati H, Akbari Sari A, Rashidian A, Mohammad K: **Population-Based Preference Weights for the EQ-5D Health States Using the Visual Analogue Scale (VAS) in Iran**. *Iranian Red Crescent medical journal* 2016, **18**(2):e21584.

19. Furlong W, Rae C, Feeny D, Gelber RD, Laverdiere C, Michon B, Silverman L, Sallan S, Barr R: **Health-Related Quality of Life Among Children with Acute Lymphoblastic Leukemia**. *Pediatric blood & cancer* 2012, **59**(4):717-724.

20. Darba J, Restovic G, Kaskens L, Balbona MA, Carbonell A, Cavero P, Jordana M, Prieto C, Molina A, Padro I: **Patient preferences for osteoporosis in Spain: a discrete choice experiment**. *Osteoporosis international : a journal established as result of cooperation*

*between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA* 2011, **22**(6):1947-1954.

21. de Bekker-Grob EW, Essink-Bot ML, Meerding WJ, Koes BW, Steyerberg EW: **Preferences of GPs and patients for preventive osteoporosis drug treatment: a discrete-choice experiment**. *Pharmacoeconomics* 2009, **27**(3):211-219.

22. Weiss TW, Gold DT, Silverman SL, McHorney CA: **An evaluation of patient preferences for osteoporosis medication attributes: results from the PREFER-US study**. *Curr Med Res Opin* 2006, **22**(5):949-960.

23. Duarte JW, Bolge SC, Sen SS: **An evaluation of patients' preferences for osteoporosis medications and their attributes: the PREFER-International study**. *Clinical therapeutics* 2007, **29**(3):488-503.

24. Scott A: **Eliciting GPs' preferences for pecuniary and non-pecuniary job characteristics**. *J Health Econ* 2001, **20**(3):329-347.

25. Penn-Kekana L, Blaauw D, Tint K, Monareng D, Chege J: **Nursing staff dynamics and implications for maternal health provision in public health facilities in the context of HIV/AIDS**. *Frontiers in Reproductive Health* 2005.

26. Gosden T, Bowler I, Sutton M: **How do general practitioners choose their practice? Preferences for practice and job characteristics**. *Journal of health services research & policy* 2000, **5**(4):208-213.

27. Wordsworth S, Skatun D, Scott A, French F: **Preferences for general practice jobs: a survey of principals and sessional GPs**. *The British journal of general practice : the journal of the Royal College of General Practitioners* 2004, **54**(507):740-746.

28. Ubach C, Scott A, French F, Awramenko M, Needham G: **What do hospital consultants value about their jobs? A discrete choice experiment**. *BMJ (Clinical research ed)* 2003, **326**(7404):1432.

29. Lagarde M, Blaauw D: **A review of the application and contribution of discrete choice experiments to inform human resources policy interventions**. *Hum Resour Health* 2009, **7**(62):10.1186.

30. Koopmanschap MA, Stolk EA, Koolman X: **Dear policy maker: have you made up your mind? A discrete choice experiment among policy makers and other health professionals**. *International journal of technology assessment in health care* 2010, **26**(2):198-204.

31. Muhlbacher AC, Nubling M: **Analysis of patients' preferences: direct assessment and discrete-choice experiment in therapy of adults with attention-deficit hyperactivity disorder**. *Patient* 2010, **3**(4):285-294.

32. Guimaraes C, Marra CA, Gill S, Simpson S, Meneilly G, Queiroz RH, Lynd LD: **A discrete choice experiment evaluation of patients' preferences for different risk, benefit, and delivery attributes of insulin therapy for diabetes management**. *Patient Prefer Adherence* 2010, **4**:433-440.

33. Nichols E, O'Hara NN, Degani Y, Sprague SA, Adachi JD, Bhandari M, Holick MF, Connelly DW, Slobogean GP: **Patient preferences for nutritional supplementation to improve fracture healing: a discrete choice experiment**. *BMJ open* 2018, **8**(4):e019685.

34. Ng-Mak D, Poon JL, Roberts L, Kleinman L, Revicki DA, Rajagopalan K: **Patient preferences for important attributes of bipolar depression treatments: a discrete choice experiment**. *Patient Prefer Adherence* 2018, **12**:35-44.

35. Herrmann A, Sanson-Fisher R, Hall A, Wall L, Zdenkowski N, Waller A: **A discrete choice experiment to assess cancer patients' preferences for when and how to make treatment decisions**. *Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer* 2018, **26**(4):1215-1220.

36. Vass CM, Rigby D, Payne K: **Investigating the Heterogeneity in Women's Preferences for Breast Screening: Does the Communication of Risk Matter?** *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2018, **21**(2):219-228.

37. Becker F, Anokye N, de Bekker-Grob EW, Higgins A, Relton C, Strong M, Fox-Rushby J: **Women's preferences for alternative financial incentive schemes for breastfeeding: A discrete choice experiment**. *PloS one* 2018, **13**(4):e0194231.

38. Porteous T, Ryan M, Bond C, Watson M, Watson V: **Managing Minor Ailments; The Public's Preferences for Attributes of Community Pharmacies. A Discrete Choice Experiment**. *PloS one* 2016, **11**(3):e0152257.

39. Kleij KS, Tangermann U, Amelung VE, Krauth C: **Patients' preferences for primary health care - a systematic literature review of discrete choice experiments**. *BMC Health Serv Res* 2017, **17**(1):476.

40. Bien DR, Danner M, Vennedey V, Civello D, Evers SM, Hiligsmann M: **Patients' Preferences for Outcome, Process and Cost Attributes in Cancer Treatment: A Systematic Review of Discrete Choice Experiments**. *Patient* 2017, **10**(5):553-565.

41. Whitty JA, Oliveira Goncalves AS: **A Systematic Review Comparing the Acceptability, Validity and Concordance of Discrete Choice Experiments and Best-Worst Scaling for Eliciting Preferences in Healthcare**. *Patient* 2018, **11**(3):301-317.

42. Harrison M, Milbers K, Hudson M, Bansback N: **Do patients and health care providers have discordant preferences about which aspects of treatments matter most? Evidence from a systematic review of discrete choice experiments**. *BMJ open* 2017, **7**(5):e014719.

43. Mansfield C, Tangka FK, Ekwueme DU, Smith JL, Guy GP, Jr., Li C, Hauber AB: **Stated Preference for Cancer Screening: A Systematic Review of the Literature, 1990-2013**. *Preventing chronic disease* 2016, **13**:E27.

44. de Bekker-Grob EW, Ryan M, Gerard K: **Discrete choice experiments in health economics: a review of the literature**. *Health economics* 2012, **21**(2):145-172.

45. Hauber AB, González JM, Groothuis-Oudshoorn CGM, Prior T, Marshall DA, Cunningham C, Ijzerman MJ, Bridges JFP: **Statistical Methods for the Analysis of Discrete Choice Experiments: A Report of the ISPOR Conjoint Analysis Good Research Practices Task Force**. *Value in Health* 2016, **19**(4):300-315.

46. Reed Johnson F, Lancsar E, Marshall D, Kilambi V, Mühlbacher A, Regier DA, Bresnahan BW, Kanninen B, Bridges JFP: **Constructing Experimental Designs for Discrete-Choice Experiments: Report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force**. *Value in Health* 2013, **16**(1):3-13.

47. Ghijben P, Lancsar E, Zavarsek S: **Preferences for oral anticoagulants in atrial fibrillation: a best-best discrete choice experiment**. *Pharmacoeconomics* 2014, **32**(11):1115-1127.

48. 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):360-375.

49. Johnson FR, Lancsar E, Marshall D, Kilambi V, Mühlbacher A, Regier DA, Bresnahan BW, Kanninen B, Bridges JF: **Constructing experimental designs for discrete-choice experiments: report of the ISPOR conjoint analysis experimental design good research practices task force**. *Value in Health* 2013, **16**(1):3-13.

50. 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.

51. Flynn TN, Louviere JJ, Peters TJ, Coast J: **Using discrete choice experiments to understand preferences for quality of life. Variance-scale heterogeneity matters**. *Social science & medicine (1982)* 2010, **70**(12):1957-1965.

52. Marshall D, Bridges JF, Hauber B, Cameron R, Donnalley L, Fyie K, Johnson FR: **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**(4):249-256.

53. Mandeville KL, Lagarde M, Hanson K: **The use of discrete choice experiments to inform health workforce policy: a systematic review**. *BMC Health Services Research* 2014, **14**(1):367.

54. de Bekker-Grob EW, Ryan M, Gerard K: **Discrete choice experiments in health economics: a review of the literature**. *Health Economics* 2012, **21**(2):145-172.

55. Ryan M, Scott DA, Reeves C, Bate A, van Teijlingen ER, Russell EM, Napper M, Robb CM: **Eliciting public preferences for healthcare: a systematic review of techniques**. *Health Technol Assess* 2001, **5**(5):1-186.

56. Spinks J, Chaboyer W, Bucknall T, Tobiano G, Whitty JA: **Patient and nurse preferences for nurse handover-using preferences to inform policy: a discrete choice experiment protocol**. *BMJ open* 2015, **5**(11):2015-008941.

57. van de Schoot T, Pavlova M, Atanasova E, Groot W: **Preferences of Bulgarian consumers for quality, access and price attributes of healthcare services-result of a discrete choice experiment**. *Int J Health Plann Manage* 2015, **18**(10).

58. Baji P, Gulacsi L, Lovasz BD, Golovics PA, Brodszky V, Pentek M, Rencz F, Lakatos PL: **Treatment preferences of originator versus biosimilar drugs in Crohn's disease; discrete choice experiment among gastroenterologists**. *Scand J Gastroenterol* 2016, **51**(1):22-27.

59. Veldwijk J, Lambooij MS, van Til JA, Groothuis-Oudshoorn CG, Smit HA, de Wit GA: **Words or graphics to present a Discrete Choice Experiment: Does it matter?** *Patient Educ Couns* 2015, **98**(11):1376-1384.

60. McCaffrey N, Gill L, Kaambwa B, Cameron ID, Patterson J, Crotty M, Ratcliffe J: **Important features of home-based support services for older Australians and their informal carers**. *Health Soc Care Community* 2015, **23**(6):654-664.

61. Veldwijk J, van der Heide I, Rademakers J, Schuit AJ, de Wit GA, Uiters E, Lambooij MS: **Preferences for Vaccination: Does Health Literacy Make a Difference?** *Medical decision making : an international journal of the Society for Medical Decision Making* 2015, **35**(8):948-958.

62. Bottger B, Thate-Waschke IM, Bauersachs R, Kohlmann T, Wilke T: **Preferences for anticoagulation therapy in atrial fibrillation: the patients' view**. *J Thromb Thrombolysis* 2015, **40**(4):406-415.

63. Adams J, Bateman B, Becker F, Cresswell T, Flynn D, McNaughton R, Oluboyede Y, Robalino S, Ternent L, Sood BG *et al*: **Effectiveness and acceptability of parental financial incentives and quasi-mandatory schemes for increasing uptake of vaccinations in preschool children: systematic review, qualitative study and discrete choice experiment**. *Health Technol Assess* 2015, **19**(94):1-176.

64. Brett Hauber A, Nguyen H, Posner J, Kalsekar I, Ruggles J: **A discrete-choice experiment to quantify patient preferences for frequency of glucagon-like peptide-1 receptor agonist injections in the treatment of type 2 diabetes**. *Curr Med Res Opin* 2015, **7**:1-32.

65. Ryan M, Gerard K: **Using discrete choice experiments to value health care programmes: current practice and future research reflections**. *Applied Health Economics and Health Policy* 2003, **2**(1):55-64.

66. Ryan M: **Discrete choice experiments in health care: NICE should consider using them for patient centred evaluations of technologies**. *BMJ (Clinical research ed)* 2004, **328**(7436):360-361.

67. Cunningham CE, Henderson J, Niccols A, Dobbins M, Sword W, Chen Y, Mielko S, Milligan K, Lipman E, Thabane L *et al*: **Preferences for evidence-based practice dissemination in addiction agencies serving women: a discrete-choice conjoint experiment**. *Addiction* 2012, **107**(8):1512-1524.

68. Ryan M: **Discrete choice experiments in health care**. *BMJ (Clinical research ed)* 2004, **328**(7436):360-361.

69. Ryan M, Bate A, Eastmond C, Ludbrook A: **Use of discrete choice experiments to elicit preferences**. *BMJ Quality & Safety* 2001, **10**(suppl 1):i55-i60.

70. Toubia O, Hauser JR: **On Managerially Efficient Experimental Designs**. *Marketing Science* 2007, **26**(6):851-858.

71. Kessels R, Goos P, Vandebroek M: **A comparison of criteria to design efficient choice experiments**. *Journal of Marketing Research* 2006, **43**(3):409-419.

72. Lourenço-Gomes L, Pinto LMC, Rebelo J: **USING CHOICE EXPERIMENTS TO VALUE A WORLD CULTURAL HERITAGE SITE: REFLECTIONS ON THE EXPERIMENTAL DESIGN**. *Journal of Applied Economics* 2013, **16**(2):303-332.

73. Carlsson F, Martinsson P: **Design techniques for stated preference methods in health economics**. *Health economics* 2003, **12**(4):281-294.

74. Sandor Z, Wedel M: **Profile Construction in Experimental Choice Designs for Mixed Logit Models**. *Marketing Science* 2002, **21**(4):455-475.

75. Arora N, Huber J: **Improving Parameter Estimates and Model Prediction by Aggregate Customization in Choice Experiments**. *Journal of Consumer Research* 2001, **28**(2):273-283.

76. Burton A, Altman DG, Royston P, Holder RL: **The design of simulation studies in medical statistics**. *Statistics in medicine* 2006, **25**(24):4279-4292.

77. Burgess L, Street DJ: **Optimal designs for choice experiments with asymmetric attributes**. *Journal of Statistical Planning and Inference* 2005, **134**(1):288-301.

78. Kanninen BJ: **Optimal design for multinomial choice experiments**. *Journal of Marketing Research* 2002, **39**(2):214-227.

79. Street DJ, Burgess L: **Optimal and near-optimal pairs for the estimation of effects in 2-level choice experiments**. *Journal of Statistical Planning and Inference* 2004, **118**(1):185-199.

80. Rose JM, Scarpa R: **Designs efficiency for non-market valuation with choice modelling: how to measure it, what to report and why**. *Australian Journal of Agricultural & Resource Economics* 2007, **52**(3):253-282.

81. Sandor Z, Wedel M: **Designing conjoint choice experiments using managers' prior beliefs**. *Journal of Marketing Research* 2001, **38**(4):430-444.

82. Ferrini S, Scarpa R: **Designs with a priori information for nonmarket valuation with choice experiments: A Monte Carlo study**. *Journal of environmental economics and management* 2007, **53**(3):342-363.

83. Sándor Z, Wedel M: **Heterogeneous conjoint choice designs**. *Journal of Marketing Research* 2005, **42**(2):210-218.

84. Sándor Z, Wedel M: **Profile construction in experimental choice designs for mixed logit models**. *Marketing Science* 2002, **21**(4):455-475.

85. Hess S, Smith C, Falzarano S, Stubits J: **Measuring the effects of different experimental designs and survey administration methods using an Atlanta managed lanes stated preference survey**. *Transportation Research Record* 2008, **2049**:144-152.

86. de Bekker-Grob EW, Donkers B, Jonker MF, Stolk EA: **Sample Size Requirements for Discrete-Choice Experiments in Healthcare: a Practical Guide**. *The Patient-Patient-Centered Outcomes Research* 2015:1-12.

87. Huber J, Zwerina K: **The importance of utility balance in efficient choice designs**. *Journal of Marketing Research* 1996:307-317.

88. Li W, Nachtsheim CJ, Wang K, Reul R, Albrecht M: **Conjoint Analysis and Discrete Choice Experiments for Quality Improvement**. *Journal of Quality Technology* 2013, **45**(1):74.

89. Girardi SN, Carvalho CL, Maas LW, Araujo JF, Massote AW, Stralen A, Souza OA: **[Preferences for work in primary care among medical students in Minas Gerais State, Brazil: evidence from a discrete choice experiment]**. *Cadernos de saude publica* 2017, **33**(8):e00075316.

90. de Vries ST, de Vries FM, Dekker T, Haaijer-Ruskamp FM, de Zeeuw D, Ranchor AV, Denig P: **The Role of Patients' Age on Their Preferences for Choosing Additional Blood Pressure-Lowering Drugs: A Discrete Choice Experiment in Patients with Diabetes**. *PloS one* 2015, **10**(10):e0139755.

91. O'Hara NN, Slobogean GP, Mohammadi T, Marra CA, Vicente MR, Khakban A, McKee MD: **Are patients willing to pay for total shoulder arthroplasty? Evidence from a discrete choice experiment**. *Canadian journal of surgery Journal canadien de chirurgie* 2016, **59**(2):107-112.

92. Norman R, Viney R, Aaronson NK, Brazier JE, Cella D, Costa DS, Fayers PM, Kemmler G, Peacock S, Pickard AS *et al*: **Using a discrete choice experiment to value the QLU-C10D: feasibility and sensitivity to presentation format**. *Qual Life Res* 2016, **25**(3):637-649.

93. Herzberg AM, Cox DR: **Recent Work on the Design of Experiments: A Bibliography and a Review**. *Journal of the Royal Statistical Society Series A (General)* 1969, **132**(1):29-67.

Contemporary Clinical Trials Communications 10 (2018) 17–28

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

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

# Investigating the impact of design characteristics on statistical efficiency within discrete choice experiments: A systematic survey

Thuva Vanniyasingam[a,b,*], Caitlin Daly[a], Xuejing Jin[a], Yuan Zhang[a], Gary Foster[a,b], Charles Cunningham[c], Lehana Thabane[a,b,d,e,f]

[a] Department of Health Research Methods, Impact, and Evidence, McMaster University, Hamilton, ON, Canada
[b] Biostatistics Unit, Father Sean O'Sullivan Research Centre, St. Joseph's Healthcare, Hamilton, ON, Canada
[c] Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, ON, Canada
[d] Departments of Paediatrics and Anaesthesia, McMaster University, Hamilton, ON, Canada
[e] Centre for Evaluation of Medicine, St. Joseph's Healthcare, Hamilton, ON, Canada
[f] Population Health Research Institute, Hamilton Health Sciences, Hamilton, ON, Canada

ARTICLE INFO

*Keywords:*
Discrete choice experiment
Systematic survey
Statistical efficiency
Relative D-efficiency
Relative D-error

ABSTRACT

*Objectives:* This study reviews simulation studies of discrete choice experiments to determine (i) how survey design features affect statistical efficiency, (ii) and to appraise their reporting quality.
*Outcomes:* Statistical efficiency was measured using relative design (D-) efficiency, D-optimality, or D-error.
*Methods:* For this systematic survey, we searched Journal Storage (JSTOR), Since Direct, PubMed, and OVID which included a search within EMBASE. Searches were conducted up to year 2016 for simulation studies investigating the impact of DCE design features on statistical efficiency. Studies were screened and data were extracted independently and in duplicate. Results for each included study were summarized by design characteristic. Previously developed criteria for reporting quality of simulation studies were also adapted and applied to each included study.
*Results:* Of 371 potentially relevant studies, 9 were found to be eligible, with several varying in study objectives. Statistical efficiency improved when increasing the number of choice tasks or alternatives; decreasing the number of attributes, attribute levels; using an unrestricted continuous "manipulator" attribute; using model-based approaches with covariates incorporating response behaviour; using sampling approaches that incorporate previous knowledge of response behaviour; incorporating heterogeneity in a model-based design; correctly specifying Bayesian priors; minimizing parameter prior variances; and using an appropriate method to create the DCE design for the research question. The simulation studies performed well in terms of reporting quality. Improvement is needed in regards to clearly specifying study objectives, number of failures, random number generators, starting seeds, and the software used.
*Conclusion:* These results identify the best approaches to structure a DCE. An investigator can manipulate design characteristics to help reduce response burden and increase statistical efficiency. Since studies varied in their objectives, conclusions were made on several design characteristics, however, the validity of each conclusion was limited. Further research should be conducted to explore all conclusions in various design settings and scenarios. Additional reviews to explore other statistical efficiency outcomes and databases can also be performed to enhance the conclusions identified from this review.

## 1. Introduction

Discrete choice experiments (DCEs) are now being used as a tool in health research to elicit participant preferences for a health product or service. Several DCEs have emerged within the health literature using various design approaches [3–6]. Ghijben and colleagues conducted a DCE to understand how patients value and trade-off key characteristics of oral anticoagulants [7]. They examined patient preferences to determine which of seven attributes of warfarin and other anticoagulants (dabigatran, rivaroxaban, apixaban) in atrial fibrillation were most important to patients [7]. With seven attributes, each with different levels, several possible combinations could be created to describe an anticoagulant. Like many DCEs, they used a fractional factorial design, a sample of all possible combinations, to create a survey with 16

* Corresponding author. Department of Clinical Epidemiology and Biostatistics, HSC 2C, McMaster University, Hamilton, ON, L8S 4L8.
*E-mail address:* thuva.vanni@gmail.com (T. Vanniyasingam).

https://doi.org/10.1016/j.conctc.2018.01.002
Received 15 August 2017; Received in revised form 1 December 2017; Accepted 8 January 2018
Available online 10 January 2018

questions that each presented three alternatives for patients to choose from. As patients selected their most preferred and second most preferred alternatives, investigators were able to model their responses to determine which anticoagulant attributes were more favourable than others. Since only a fraction of combinations are typically used in DCEs, it is important to use a statistical efficiency measure to determine how well the fraction represents all possible combinations of attributes and attribute levels.

There is no single specific design to yield optimal results of a discrete choice experiment (DCE). They can vary in their level of statistical efficiency and response burden. The variation in designs can be seen in several reviews covering various decades [8–13]. While presenting all possible combinations of attributes and attribute levels will always yield 100% statistical efficiency, this is not feasible in many cases. For fractional factorial designs, a statistical efficiency measure can be used to reduce the bias of the fraction selected. A common measure to assess statistical efficiency of these partial designs is relative design efficiency (D-efficiency) [14,15]. For a design matrix $X$, the formula is as follows:

$$Relative\ D-efficiency = 100*\frac{1}{N_D\left|(X'X)^{-1}\right|^{1/p}},$$

where $X'X$ is the information matrix, $p$ is the number of parameters, and $N_D$ is the number of rows in the design [16].

To yield a statistically efficient design, a design will be orthogonal and balanced, or nearly so. A design is balanced when attribute levels are evenly distributed [17]. This occurs when the off-diagonal elements in the intercept row and column are zero [16]. It is orthogonal when the pairs of attribute levels are evenly distributed [17], that is when the submatrix *of X'X*, without the intercept, is a diagonal matrix [16]. Therefore, to maximize relative D-efficiency, we need to reduce the $(X'X)^{-1}$ matrix to a diagonal that equals $\frac{1}{N_d}I$ for a suitably coded $X$ [16].

Relative D-efficiency is often referred to as relative design optimality (D-optimality) or is described using its inverse, design error (D-error) [18]. It ranges from 0% to 100%, where 100% indicates a statistically efficient design. A measure of 100% can still be achieved with fractional factorial designs; however, there is limited knowledge as to how the various design characteristics impact statistical efficiency.

Identifying the impact of DCE design characteristics on statistical efficiency will bring more power to investigators, particularly research practitioners, during the design stage. They can reduce the variance of estimates by manipulating their designs to construct a simpler DCE that is statistically efficient and minimizes participants' response burden. Currently there are several studies exploring DCE designs. These studies range from comparing or introducing new statistical optimality criteria [19,20] to approaches for generating DCEs [14] to exploring the impact of different prior specifications on parameter estimates [21–23]. To our knowledge, the results of these findings have not been summarized. This may be due to the variation in objectives and outcomes across studies, making it hard to synthesize information and draw conclusions. As part of a previous simulation study, a literature review was also performed to report the DCE design characteristics explored by investigators in simulation studies [1]. However, information on the impact of these design characteristics on relative D-efficiency, the common outcome among each study, was not assessed.

The primary aim of this systematic survey was to review simulation studies to determine design features that affect the statistical efficiency of DCEs—measured using relative D-efficiency, relative D-optimality, or D-error; and to appraise the completeness of reporting of the studies using the criteria for reporting simulation studies [24].

## 2. Methods

### 2.1. Eligibility criteria

The inclusion criteria were comprised of simulation studies of DCEs that explored the impact of DCE design characteristics on relative D-efficiency, D-optimality, or D-error. Search terms were first searched by variations in spelling and acronyms of individual terms and then combined into one search. Studies were restricted to English articles. Studies were excluded if they were not related to DCEs (or not referred to as stated preference, latent class, and conjoint analysis), were applications of DCEs, empirical comparisons, reviews or discussions of DCEs, or simulation studies that did not explore the impact of DCE design characteristics on statistical efficiency. Duplicate publications, meeting abstracts, letter, commentary, editorials and protocols, books and pamphlets were also excluded.

### 2.2. Search strategy

Two rounds of electronic searches were conducted covering the period from inception to Sept 19, 2016. The first round was performed on all databases from inception to July 20, 2016. The second round extended the search until Sept 19, 2016. The databases searched were Journal Storage (JSTOR), Science Direct, PubMed, and OVID which included a search within EMBASE. Studies identified from Vanniyasingam et al.'s literature review, that were not identified in this search, were also considered [1]. Table 1 (Supplementary Files) presents the detailed search strategy for each database.

### 2.3. Study selection

Four reviewers worked independently and in duplicate to screen titles and abstracts of all citations identified in the search. Any potentially eligible article identified by either reviewer, from each pair, proceeded to the full-text review stage. The same authors then, independently and in duplicate, applied the above eligibility criteria to screen the full text of these studies. Disagreement regarding eligibility were resolved through discussion. If a disagreement was unresolved, a third author (a statistician) adjudicated and resolved the conflict. After full-text screening forms were consolidated amongst pairs, data was extracted from eligible studies. Both the full-text screening and data extraction forms were first piloted with calibration exercises to ensure consistency in reviewer reporting.

### 2.4. Data extraction process

A Microsoft Excel spreadsheet was used to extract information related to general study characteristics, DCE design characteristics that varied or were held fixed, and the impact of the varied design characteristics on statistical efficiency.

### 2.5. Reporting quality

The quality of reporting was also assessed by extracting information related to the reporting guidelines for simulation studies described by Burton and colleagues [24]. Some components were modified to be more tailored for simulation studies of DCEs. This checklist included whether studies reported:

- A detailed protocol of all aspects of the simulation study
- Clearly defined aims

- The number of failures during the simulation process (defined as the number of times it was not possible to create a design given the design component restrictions)
- Software used to perform simulations
- The random number generator or starting seed, the method(s) for generating DCE datasets
- The scenarios to be investigated (defined as the specifications of each design characteristic explored and overall total number of designs created)
- Methods for evaluating each scenario
- The distribution used to simulate the data (defined as whether or not the design characteristics explored are motivated by real-world or simulation studies)
- The presentation of the simulation results (defined as whether authors used separate subheadings for objectives, methods, results and discussion to assist with clarity). Information presented in graphs or tabular form but not written as detailed in the manuscripts were counted for if they were presented in a clear and concise manner.

One item was added to the criteria to determine whether or not studies provided a rationale for creating the different designs. Reporting items excluded were: a detailed protocol of all aspects of the simulation study, level of dependence between simulated designs, estimates to be stored for each simulation, summary measures to be calculated over all simulations, and criteria to evaluate the performance of statistical methods (bias, accuracy, and coverage). We decided against checking whether a detailed protocol was reported because the studies of interest were focussed on only the creation of DCE designs. The original reporting checklist is tailored towards randomized controlled trials or prognostic factor studies with complex situations seen in practice [24]. The remaining items were excluded because the specific statistical efficiency measures were required for studies to be included in the study. Also, there were no summary measures to be calculated over all simulations, and no results to measure bias, accuracy and coverage.

When studies referred to supplementary materials, these materials were also reviewed for data extraction.

Three of the four reviewers, working in pairs, performed data abstraction independently and in duplicate. Pairs resolved disagreements through discussion or, if necessary, with assistance from another statistician.

## 3. Data analysis

### 3.1. Agreement

Agreement between reviewers on the studies' eligibility based on full text screening was assessed using an unweighted kappa. A kappa value was indicative of poor agreement if it was less than 0.00, slight agreement if it ranged from 0.00 to 0.20, fair agreement between 0.21 and 0.40, moderate agreement between 0.41 and 0.60, substantial agreement between 0.61 and 0.80, and almost perfect agreement when greater than 0.80 [25].

### 3.2. Data synthesis and analysis

The simulation studies were assessed by the details of their DCE designs. More specifically, the design characteristics investigated and their ranges were recorded along with their impact on statistical efficiency (relative D-efficiency, D-optimality, or D-error). Adherence to reporting guidelines was also recorded [24].

## 4. Results

### 4.1. Search strategy and screening

A total of 371 papers were identified from the search and six were selected from a previous literature search that used snowball sampling [1]. From this, 43 were removed as duplicates and 245 were excluded during title and abstract screening. Of the remaining 77 studies for full text screening, three needed to be ordered [26–28] and one we were unable to obtain a full text for [29], 18 did not relate to DCEs (or include terms such as discrete choice, DCE, choice-based, binary choice, stated preference, latent class, conjoint analysis, or fractional factorial design, factorial design); 17 did not perform a simulation analysis; 1 did not use its simulations to create DCE designs; 22 did not assess the statistical efficiency of designs using relative D-optimality, D-efficiency, or D-error measures; 4 did not compare the impact of various design characteristics on relative D-efficiency or D-optimality or D-error; and 1 was not a peer-reviewed manuscript. Details of the search and screening process are presented in a flow chart in Fig. 1(Appendix).

Finally, nine studies remained after full-text screening. The unweighted kappa for measuring agreement between reviewers on full text eligibility was 0.53, indicating a moderate agreement [25]. Of the 9 studies included, 1 was published in Marketing Science, 1 in the Journal of Statistical Planning and Inference, 2 in the Journal of Marketing Research, 1 in the International Journal of Research in Marketing, 2 in Computational Statistics and Data Analysis, 1 in BMJ Open, and 1 in Transportation Research Part B: Methodological.

The number of statistical efficiency measures, scenarios, and design characteristics varied from study to study. Of the outcomes assessed for each scenario, four studies reported relative D-efficiency [1,30–32], two D-error [33,34], three $D_b$-error (a Bayesian variation of D-error) [30,35,36], and two percentage changes in D-error [34,37]. Of the design characteristics explored, one study explored the impact of attributes on statistical efficiency [1], two explored alternatives [1,30], one explored choice tasks [1], two explored attribute levels [1,32], two explored choice behaviour [33,37], three explored priors [30,31,34], and four explored methods to create the design [30,34–36]. Results are further described below based for each design characteristic. Details of the ranges of each design characteristic investigated and corresponding studies are described in Table 2 (Supplementary Files).

### 4.2. Survey-specific components

The simulation studies had several conclusions based on the number of choice tasks, attributes, and attribute levels; the type of attributes (qualitative and quantitative); and the number of alternatives. First, ***increasing the number of choice tasks*** increased relative D-efficiency (or improved statistical efficiency) across several designs with varying numbers of attributes, attribute levels, and alternatives [1]. Second, ***increasing the number of attributes*** generally (i.e. not monotonically) decreased relative D-efficiency. For designs with a large number of attributes and a small number of alternatives per choice task, a DCE could not be created [1]. Third, ***increasing the number of levels*** **within** ***attributes*** (from 2 to 5) decreased relative D-efficiency. In fact, binary attribute designs had higher statistical efficiency in comparison to all other designs with varying numbers of alternatives (2–5), attributes (2–20), and choice tasks (2–20). However, higher relative D-efficiency measures were also found when the number of attribute levels equalled the number of alternative [1]. Fourth, **increasing the number of alternatives** improved statistical efficiency [1,30]. Fifth, for ***designs with only binary attributes and one quantitative (continuous) attribute, it***

***was possible to create locally optimal designs.*** To further clarify this result, DCEs were created where two of three alternatives were identical or differed only by an unrestricted continuous attribute *(e.g. size, weight, or speed)*. The third alternative differed from the two others in the binary attributes [32]. The continuous variable was unrestricted and used as a "manipulating" attribute to offset dominating alternatives or alternatives with a zero probability of being selected in a choice task. This finding, however, was conditional on the type of quantitative variable and was concluded to be unrealistic in the study [32]. Details of the studies exploring these design characteristics are presented in Tables 1a and 1b (Appendix).

### 4.3. Incorporating choice behaviour

Two approaches were used to incorporate response behaviour when designing a DCE. First, the order of the statistical efficiency of designs from highest to lowest were if they: (i) incorporated covariates relating to response behaviour, (ii) incorporated covariates not relating to response behaviour, and (iii) did not incorporate any covariates. Second, among binary choice designs, stratified sampling strategies had higher statistical efficiency measures in comparison to randomly sampled strategies. This was most apparent when stratification was performed on both expected choice behaviour (e.g. 2.5% of the population selects Y = 1, remaining selects Y = 0) and on a binary independent factor associated with the response behaviour. Similar efficiency measures were found when there was an even distribution (50% of the population selects Y = 1) across approaches [37] (Table 1c, Appendix).

### 4.4. Bayesian priors

Studies also explored the impact of parameter priors and heterogeneity priors. Increasing the parameter prior variances [30] or misspecifying priors (in comparison to correctly specifying priors) [34] reduced statistical efficiency. In one study, mixed logit designs that incorporated respondent heterogeneity had higher statistical efficiency measures than designs ignoring respondent heterogeneity [31]. However, misspecifying the heterogeneity prior had negative implications. In fact, underspecifying the heterogeneity prior had a greater loss in efficiency in comparison to over specifying it [31] (Table 1d, Appendix).

### 4.5. Methods to create the design

Several simulation studies compared various methods to create a DCE design against other design settings (Table 1e, Appendix). First, relative statistical efficiency measures were highest when the method to create a design matched the method used for the reference design setting [30,34,36]. For example, a multinomial logit (MNL) generated design had the highest statistical efficiency in an MNL design setting, in comparison to a cross-sectional mixed logit or a panel mixed logit design setting [36]. Similarly, a partial rank-order conjoint experiment yields highest statistical efficiency for a design setting of the same type in comparison to a best-choice experiment, best-worst experiment or orthogonal design setting [30]. Second, among frequentist (non-Bayesian) approaches, the order of designs yielding the highest statistical efficiency is d-optimal rank designs, d-optimal choice designs, near-orthogonal, random designs, and balanced overlap designs for full rank order and partial rank order choice experiments [35]. Third, a semi-Bayesian d-optimal best-worst choice design outperformed frequentist and Bayesian-derived designs, while yielding similar statistical efficiency measures as semi-Bayesian d-optimal best-worst choice designs [30].

### 4.6. Reporting of simulations studies

All studies clearly reported the primary outcome, rationale and methods for creating designs, and methods to evaluate each scenario. Reporting the objective was unclear in two studies and no study reported any failures in the simulations. In many cases, such as in Vermeulen et al.'s study [36], the distribution from which random numbers were selected from were described, however no study specified the starting seeds. Also, no study reported the number of times it was not possible to create a design given the design component restrictions except for Vanniyasingam et al. [1], who specified that designs with a larger number of attributes could not be created with a small number of alternatives or choice task. The total number of designs and the range of design characteristics explored were either written or easily identifiable from figures and tables. Five studies reported the software used for the simulation studies and one study reported the software used for only one of the approaches to create a design. Four studies chose design characteristics that were motivated by real-world scenarios or previous literature, while four were not motivated by other studies. Details of each study's reporting quality are broken down in Table 2 (Appendix).

## 5. Discussion

### 5.1. Summary of findings

Several conclusions can be drawn from the nine simulation studies included in this systematic survey of investigating the impact of design characteristics on statistical efficiency. Factors recognized for improving statistical efficiency of a DCE include (i) increasing the number of choice tasks or alternatives; (ii) decreasing the number of attributes, and levels within attributes; (iii) using model-based designs with covariates or sampling approaches that incorporate response behaviour; (iv) incorporating heterogeneity in a model-based design; (v) correctly specifying Bayesian priors and minimizing parameter prior variances; and (vi) the method to create the DCE design is appropriate for the research question and design at hand. Lastly, optimal designs could be created using 3 alternatives with all binary attributes except one continuous attribute. Here, two alternatives were identical or differed only by the continuous attribute and the third alternative differed by the binary attribute. Overall, studies were detailed in their descriptions of simulation studies. Improvement is needed to ensure the study objectives, number of failures, random number generators, starting seeds, and the software used are clearly defined.

### 5.2. Discussion of simulation studies

Many of the studies agree with the formula for relative d-efficiency, however some appear to contradict it. Conclusions related to choice tasks, alternatives, attributes, and attribute levels all agree with the relative d-efficiency formula where increasing the number of parameters (with attributes and attribute levels) will reduce statistical efficiency and increasing the number of choice tasks improves it. Also, when the number of attribute levels and alternatives are equal, increasing the number of attribute levels may compromise statistical efficiency, however it can be compensated by increasing the number of alternatives (which may increase $N_d$). A conclusion that cannot be directly deduced from the formula are in relation to designs with qualitative and unrestricted quantitative attributes. Grabhoff and colleagues were able to create optimal designs where two alternatives were either completely identical or only differed by a continuous variable [32]. With less information provided within each choice task (or more

overlaps), we expect a lower statistical efficiency measure. Their design approach first develops a design solution using the binary attributes and then adds the continuous attribute to maximize the efficiency. This was a continuation of Kanninen's study who explained that the continuous attribute could be used to offset dominating alternatives or alternatives that carried a zero probability of being selected by a respondent [38]. It acted as a function of a linear combination of the other binary attributes. This continuous attribute, however, was conditional on the type of quantitative variable (such as size). Other types (such as price) may result in the “red bus/blue bus” parody) [32].

#### 5.2.1. Importance

To our knowledge, this systematic survey is the first of its kind in synthesizing information on the impact of DCE design characteristics on statistical efficiency in simulation studies. Other studies have focussed on the reporting of applications of DCEs [39], and the details of DCEs and alternative approaches [40]. Systematic and literature reviews have highlighted the design type (e.g. fractional factorial or full factorial designs) and statistical methods used to analyze applications of DCEs within health research [2,11,13,41]. Exploration into summarizing the results of simulation studies is limited.

#### 5.2.2. Strengths

This study has several strengths. First, it focuses on simulation studies which are able to (i) explore several design settings to answer a research question in a single study that real world applications are unable to; (ii) act as an instrumental tool to aid in the understanding of statistical concepts such as relative d-efficiency; and identify patterns in design characteristics for improving statistical efficiency. Second, it appraises the rigour of the simulations performed, through evaluating the reporting quality, to ensure the selected studies are appropriately reflecting high quality DCEs. Third, it provides an overview for investigators to assess the scope of the literature for future simulation studies. Fourth, the results presented here can provide further insight for investigators on patterns that exist in statistical efficiency. For example, if some design characteristics must be fixed (such as the number of attributes and attribute levels), investigators can manipulate others (e.g. number of alternatives or choice tasks) to improve both the statistical optimality and response efficiency of the DCE.

#### 5.2.3. Limitations

There are some caveats to this systematic survey that may limit the direct transferability of these results to empirical research. First, the search for simulation studies of DCEs was only performed within health databases. Despite capturing a few studies from marketing journals in our search, we did not explore grey literature, statistics journals, or marketing journals. Second, we only describe the results for three outcomes (relative D-efficiency, D-error, and D-optimality) while some studies have reported other statistical efficiency measures. Third, with only nine included studies, each varying in objectives, it was not possible to make strong conclusions at this stage. Only summary findings of each study could be presented. Last, informant (or response) efficiency was not considered when extracting results from each simulation study. We recognize that incorporating participants' cognitive burden has a critical impact on the effect of overall estimation precision[42]. Integrating response efficiency with statistical efficiency would refine the focus on the structure, content, and pretesting of the survey instrument itself.

#### 5.2.4. Further research

This systematic survey provides many avenues for further research. First, these results can be used as hypotheses for future simulation studies to test and compare in various DCE scenarios. Second, a review can be performed on other statistical efficiency outcomes such as the precision of parameter estimates or reduction in sample size to compare the impact of each design characteristic. Third, a larger review should be conducted to explore simulation studies within economic, marketing, and pharmacoeconomic databases.

## 6. Conclusions

Presenting as many possible combinations (via choice tasks or alternatives) or decreasing the total number of all possible combinations (via attributes or attribute levels) will improve statistical efficiency. Model-based approaches were popularly used to create designs. These models varied from adjusting for heterogeneity, including covariates, and using a Bayesian approach. They were also applied to several different design settings. Overall reporting was clear, however improvements can be made to ensure the study objectives, number of failures, random number generators, starting seeds, and the software used are clearly defined. Further areas of research to aid in solidifying the conclusions from this paper include a systematic survey of other outcomes related to statistical efficiency, a survey on databases outside of health research that also use DCEs, and a large-scale simulation study to test each conclusion from these simulation studies.

## Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

## Competing interests

TV, CD, XJ, YZ, GF, and LT declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work. CC's participation was supported by the Jack Laidlaw Chair in Patient-Centered Health Care.

## Data sharing

As this is a systematic survey of simulation studies no data of patients exist. Data from each simulation study needs to be obtained directly from their corresponding authors.

## Author contributions

All authors provided intellectual content for the manuscript and approved the final draft.

TV contributed to the conception and design of the study; screened and extracted data; performed the statistical analyses; drafted the manuscript; approved the final manuscript; and agrees to be accountable for all aspects of the work in relation to accuracy or integrity.

CD, XJ, and YJ assisted in screening and extracting data; critically assessed the manuscript for important intellectual content; approved the final manuscript; and agrees to be accountable for all aspects of the work in relation to accuracy or integrity.

LT contributed to the conception and design of the study; provided statistical and methodological support in interpreting results and drafting the manuscript; approved the final manuscript; and agrees to be accountable for all aspects of the work in relation to accuracy or integrity.

CC and GF contributed to the design of the study; critically assessed the manuscript for important intellectual content; approved the final manuscript; and agree to be accountable for all aspects of the work in relation to accuracy or integrity.

**Appendix**

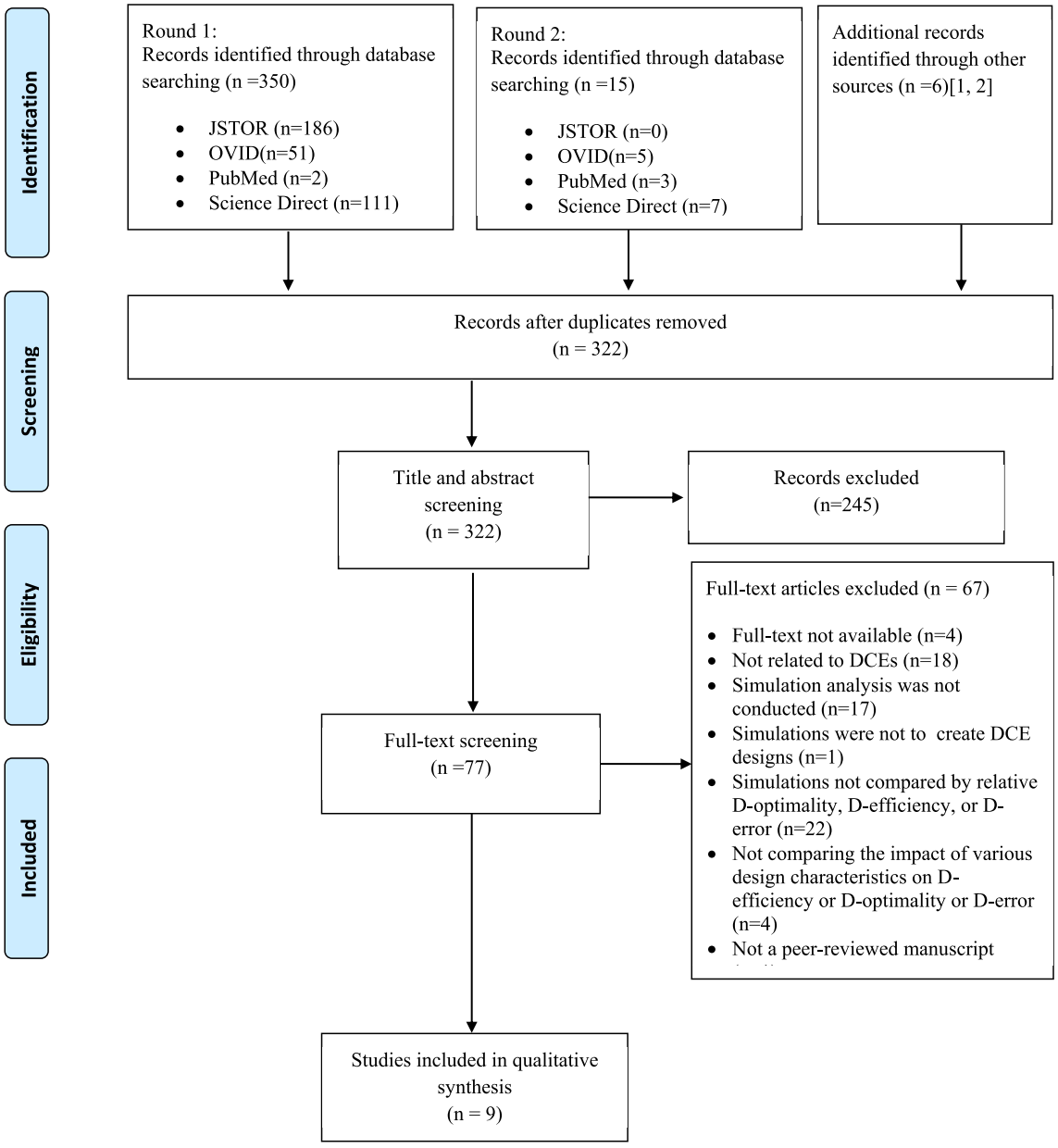



**Fig. 1.** PRISMA flow diagram.

Contemporary Clinical Trials Communications 10 (2018) 17–28

Table 1a
Studies investigating the number of choice tasks, attributes, and attribute levels

| Author, Year | Outcome of interest | Method to create design | Design setting | Distribution of Priors of parameter estimates | Choice sets | Alternatives | Attributes | Attribute levels | Results |
|---|---|---|---|---|---|---|---|---|---|
| # Choice tasks | | | | | | | | | |
| Vanniyasingam, 2016 [1] | Relative D-efficiency | . | . | no priors | 2–20 | 2–5 | 2–20 | 2–5 | 1) Generally, as the number of choice tasks increases, relative D-efficiency increases |
| # Attributes | | | | | | | | | |
| Vanniyasingam, 2016 [1] | Relative D-efficiency | . | . | no priors | 2–20 | 2–5 | 2–20 | 2–5 | 1) Generally, increasing# attributes, decreases relative D-efficiency (not monotonically)<br>2) designs with a small# of alternatives and large number of attributes could not be created. |
| # Attribute levels | | | | | | | | | |
| Vanniyasingam, 2016 [1] | Relative D-efficiency | . | . | no priors | 2–20 | 2–5 | 2–20 | 2–5 | 1) Generally, increasing# attribute levels, decreases relative D-efficiency<br>2) designs yield higher D-efficiency measures when the# attribute levels match the number of alternatives<br>3) Generally, binary attributes perform best across all other designs |
| Graβhoff, 2013 [32] | Efficiency | . | . | $\beta_1 = 0, \beta_2 = 1$ | . | 3 | 1–7 | Unrestricted quantitative (continuous) and qualitative (binary) attributes | 1) Design-optimality was achieved when two alternatives were identical or differed only in the (unrestricted) quantitative variable, while the third alternative varies in all of the qualitative components. |

Table 1b
Studies investigating the number of alternatives on statistical efficiency

| # Alternatives | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Author, Year | Outcome of interest | Method to create design | Design setting | Distribution of Priors of parameter estimates | Choice sets | Alternatives | Attributes | Attribute levels | Results |
| Vermeluen, 2010 [30] | $D_b$-error | 1: Best choice experiment<br>2: Partial rank-order conjoint experiment<br>3: Best-worst choice experiment | 1: Partial best choice experiment<br>2: Rank-order conjoint experiment<br>3: Best-worst experiment | Parameter estimates follow a normal distribution with mean priors<br>[-0.75 0–0.75 0–0.75 0 0.75 0.75]<br>Variance priors range:<br>a) 0.04<br>b) 0.5 | 9 | 4, 5, 6 | 5 | 3 three-level, 2 two-level attributes | Alternatives<br>1) Increasing the number of alternatives reduced the $d_b$-error across all scenarios |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Vanniyasingam, 2016 [1] | Relative D-efficiency | 4: Orthogonal designs<br>Random allocation | . | No priors | 2–20 | 2–5 | 2–20 | 2–5 | Alternatives:<br>1) Increasing# alternatives, increases relative D-efficiency<br>except for binary attributes which best performed with only 2 alternatives<br>2) designs yield higher D-efficiency measures when the# attribute levels match the number of alternatives |

Table 1c
Studies investigating the incorporation of choice behaviour on statistical efficiency

| Author, Year | Outcome of interest | Method to create design | Design setting | Sample size | Choice sets | Altern-atives | Attri-butes | Attribute levels | Results |
|---|---|---|---|---|---|---|---|---|---|
| Crabbe, 2012 [33] | Local D-error | 1) Individually adapted sequential Bayesian designs (IASB) with covariates incorporated<br>2) IASB designs, no covariates<br>3) single nearly orthogonal designs, no covariates | 1) Choice behaviour is influenced by 2 covariates<br>2) Choice behaviour is NOT influenced by 2 (irrelevant) covariates | 25, 250 | 16 | 3 | 3 | 3 | Across all design settings and sample sizes:<br>1) Despite IASB designs incorporating two *irrelevant* covariates of participant behaviour, they still more statistically efficient than designs that do not incorporate any covariates.<br>2) IASB designs with two relevant covariates perform better (in terms of D-efficiency) in comparison to IASB designs with to irrelevant covariates, holding everything else constant. |
| Donkers, 2003 [37] | Average percentage change in D-error | Design incorporates the proportion of the population selecting y = 1, which varies from 2.5%, 5%, 10%, 15%, and 50% of the population. Results of D-error compared to random sampling from population. | . | Sample selection is dependent on proportion that selects Y = 1 | | 2 | 2 | 1 binary, 1 continuous (Distribution of binary attribute when X = 1: 50% or 10% of the time) | 1) As the proportion of the population selecting Y = 1 increases from 2.5% to 50%, D-efficiency improves.<br>2) As more individuals select 1, the magnitude of the reduction in D-error decreases (in comparison to when a random sample is used). The highest reduction in D-error (or improvement in D-efficiency) is when only 2.5% of the population selects y = 1.<br>3) Above results are consistent when binary attribute (x = 1) is distributed 10% or 50% of the time within the DCE. |
| Donkers, 2003 [37] | Average percentage change in D-error | Design incorporates the proportion of the population selecting y = 1, which varies from 2.5%, 5%, 10%, 15%, and 50% of the population. Results of D-error compared to random sampling from population. | . | Sample selection is dependent on:<br>a) y only<br>b) y and x<br>c) x only | | 2 | 2 | 1 binary, 1 continuous; binary attribute is unevenly distributed with x = 1 only 10% of the time | Type of sample selection (y only, y and x, x only)<br>1) Designs with sample selection on both y and x yields higher statistical efficiency than designs with sample selection on y only or x only, where y is the outcome, and x is an attribute. |





Contemporary Clinical Trials Communications 10 (2018) 17–28

Contemporary Clinical Trials Communications 10 (2018) 17–28



Table 1d

Studies investigating Bayesian priors on statistical efficiency

| Author, Year | Yu, 2009 [31] | Vermeulen, 2010 [30] | Bliemer, 2010 [34] |
|---|---|---|---|
| Outcome of interest | Relative local D-efficiency | $D_b$-error | D-error and percentage change in D-error |
| Describe the scenario | 8 different designs, each compared within 5 different parameter spaces/design settings. | Comparing four designs within 3 settings for designs varying in alternatives and variance priors of parameters | Misspecification of prior parameter values |
| Method to create design | Models 1–3: Mixed logit semi-Bayesian d-optimal design<br>Model 4: ML locally d-optimal design<br>Models 5,6: MNL Bayesian D-optimal design<br>Model 7: MNL Locally D-optimal design<br>Model 8: Nearly orthogonal design | Model 1: Best choice experiment<br>Model 2: Partial rank-order conjoint experiment<br>Model 3: Best-worst choice experiment<br>Model 4: Orthogonal designs | Model 1: MNL<br>Model 2: Cross-sectional mixed logit<br>Model 3: Panel mixed logit |
| Design setting | Parameters were drawn from a normal distribution:<br>Mean: Models 1–3 = μ; Model 4: μ;<br>Model 5: $\mu + 0.51 \times I_8$;<br>where μ = [-0.5 0–0.5 0–0.5 0–0.5 0]'<br>Covariance: Model 1 = $0.25 \times I_8$; Model 2 = $I_8$;<br>Model 3 = $2.25 \times I_8$; Model 4 = $I_8$; Model 5 = $I_8$ | Setting 1: Partial best choice experiment<br>Setting 2: Rank-order conjoint experiment<br>Setting 3: Best-worst experiment | Setting 1: MNL model<br>Setting 2: Cross-sectional mixed logit model<br>Setting 3: Panel mixed logit model |
| Heterogeneity prior | Model 1 = $1.5 \times 18$; Model 2 = $1_8$;<br>Model 3: $0.5 \times 18$; Model 4 = $1_8$;<br>Model 5–7: $0_8$; Model 8: no prior,<br>Where $1_8$ is an 8-dimensional identity matrix | . | . |
| Distribution of Priors of parameter estimates | Model 1–3: Normal distribution with fixed mean, covariance I8<br>Model 4: fixed mean<br>Model 5: Normal distribution with fixed mean, covariance $9 \times$ I8<br>Model 6: Normal distribution with fixed mean, covariance I8<br>Model 7: fixed mean<br>Model 8: no priors<br>Where fixed mean vector is:<br>[-0.5 0–0.5 0–0.5 0–0.5 0]' | Parameter estimates follow a normal distribution with mean priors:<br>[-0.75 0–0.75 0–0.75 0 0.75 0.75],<br>and variance prior range:<br>a) 0.04<br>b) 0.5 | Settings 1–3: Assumed true value of parameters: $\beta_0 = -0.5$,<br>$\beta_1$: Normal (−0.07, 0.03),<br>$\beta_2$: Uniform (−1.1, −0.8),<br>$\beta_3$: Normal (−0.6, 0.15),<br>$\beta_4 = -0.3$<br>Designs 1–3 (from scenario 3):<br>β0 = −0.5, β1: Normal (−0.05, 0.02), β2: Uniform (−0.9, 0.2), β3: Normal (−0.8, 0.2), β4 = −0.2 |
| Choice sets | 12 | 9 | 12 |
| Alternatives | 3 | 4, 5, 6 | 2 |
| Attributes | 4 | 5 | 3 |
| Attribute levels | 3 | 3 three-level, 2 two-level attributes | 2 three-level attributes, 1 four-level attribute |
| Results | 1) Across all 5 design settings: Mixed logit model designs performed substantially better than designs that ignored respondent heterogeneity<br>2) Comparing Semi-Bayesian designs (Models 1–3):<br>a) Overspecifying the heterogeneity prior (Model 1) does not have too large of a negative impact on efficiency<br>b) Underspecifying the heterogeneity prior (Model 3) has a greater loss in efficiency in comparison to overspecifying it (Model 1)<br>3) Results remain consistent across other design setting such as: $2 \times 3 \times 4/2/24$ and $2 \times 2 \times 3/3/12$ | Parameter priors:<br>1) Increasing the parameter prior variances increased the $d_b$-error. | 1) D-errors of designs with misspecified priors were higher than designs with correctly specified priors (from scenario. |

Contemporary Clinical Trials Communications 10 (2018) 17–28



Table 1e

Studies investigating methods to create DCE designs on statistical efficiency

| Author, Year | Vermeulen, 2011 [35] | Bliemer, 2010 [34] | Vermeulen, 2008 [36] | Vermeulen, 2010 [30] | Vermeulen, 2010 [30] |
|---|---|---|---|---|---|
| Outcome | $D_b$-error | D-error | $d_b$-error | $D_b$-error | Relative D-efficiencies |
| Describe the scenario | Comparing different designs to create DCEs for 2 settings: full rank- and partial rank-order choice-based conjoint experiments | Comparing three types of designs against each other and an orthogonal design. | Comparing different designs to create DCEs in 2 settings: a presence and absence of 'no-choice' alternative in DCEs | Comparing different designs to create DCEs in 3 settings and with varying alternatives and variance priors | Comparing semi-Bayesian D-optimal best-worst design with 6 benchmark designs |
| Choice sets | 9 | 9, 12 | 16 | 9 | 9 |
| Alternatives | 4 | 2,3 | 2 and "no choice" alternative | 4,5, 6 | 4 |
| Attributes | 5 | 3,4 | 3 | 5 | 5 |
| Attribute levels | $3^3 2^2$ | $3^2 4^1$ | $3^2 2^1$ | $3^2 2^2$ | $3^2 2^2$ |
| Method to create design | Design:<br>1. Bayesian D-optimal ranking<br>2. D-optimal choice<br>3. Balanced overlap<br>4. Near-orthogonal<br>5. Random | Design:<br>1. MNL design<br>2.Cross-sectional mixed logit design (heterogeneity prior = 0), Panel mixed logit design<br>Priors: fixed parameters, priors equal to the mean | Design:<br>1. MNL model<br>2. Extended no-choice MNL<br>3. Nested no-choice MNL<br>4. Model-robust | Design:<br>1. Best choice<br>2. Partial rank-order conjoint<br>3: Best-worst choice<br>4: Orthogonal | Design:<br>1. Semi-Bayesian D-optimal best-worst choice<br>2. Utility-neutral best-worst choice<br>3. Semi-Bayesian D-optimal choice<br>4. Utility-neutral choice<br>5. Nearly orthogonal<br>6. Random<br>7. Balanced attribute level overlap |
| Design setting | Setting:<br>1. Full rank-order choice-based conjoint experiments<br>2. Partial rank-order choice-based conjoint experiments | Setting:<br>1. MNL<br>2. Cross-sectional mixed logit<br>3. Panel mixed logit model<br>4. Orthogonal (within alternatives) design | Setting:<br>1. Extended no-choice multinomial logit model<br>2: Nested no-choice multinomial logit model | Setting:<br>1. Partial best choice experiment<br>2. Rank-order conjoint experiment<br>3. Best-worst experiment | Setting:<br>Design 1 was set as the comparator design against all other designs (above#2–7) |
| Priors | | Settings 1–3: Assumed priors correspond to true parameter values:<br>Case 1:<br>$\beta_1 \sim N(0.6, 0.2)$, $\beta_2 \sim N(-0.9, 0.2)$, $\beta_3 = -0.2$, $\beta_4 = 0.8$;<br>Case 2:<br>$\beta_1 \sim U(-0.9, 0–0.5)$, $\beta_2 \sim N(-0.8, 0.2)$, $\beta_3 \sim U(-1.5,-1.0)$ | | Priors for each setting:<br>Parameter estimates follow a normal distribution with mean priors:<br>[-0.75 0–0.75 0–0.75 0 0.75 0.75] and variance priors:<br>a) 0.04<br>b) 0.5 | Coefficients come from an 8-dimensional normal distribution with<br>mean prior: [ 1.5 0 1.5 0 1.5 0 1.5 1.5] and variance prior |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | Case 3: $\beta 0 = -0.5$, $\beta_1 \sim N(-0.05, 0.02)$, $\beta_2 \sim U(-0.9, 0.2)$, $\beta_3 \sim N(-0.8, 0.2)$, $\beta_4 = -0.2$; Setting 4: Misspecification of prior parameters | | | (for every coefficient): 0.5. |
| Results | D-opt rank > D-opt. choice > Near-orthogonal > Random > Balanced overlap | Models estimated using designs specifically generated for that model outperform designs generated for different mode forms. | Models estimated using designs specifically generated for that model outperform designs generated for different mode forms.<br>For setting 1:<br>Model 2 > 4 > 3 > 1<br>For setting 2:<br>Model 3 > 4 > 2 > 1 | 1)Models estimated using designs specifically generated for that model outperform designs generated for different mode forms<br>2) Models 1,2,3 > 4 | 1. Design 1 > 2, 4, 5, 6, 7<br>2. Design 1's optimality is similar to Design 3 outstanding |

Comment: The greater than sign " > " indicates which method performed better than another method in terms of statistical efficiency.

Table 2
Reporting items of simulations studies

| Author, Year | Protocol | Primary outcome | Clear aim | Number of failures | Software | Random number generator or starting seed | Rationale for creating designs | Methods for creating designs | Scenarios: Total number of designs | Scenarios: Range of design characteristics explored | Method to evaluate each scenario | Distribution used to simulate data* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Vermeulen, 2011 [35] | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 |
| Yu, 2009 [31] | 0 | 1 | 1 | 0 | 0, 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| Bliemer, 2010 [34] | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| Crabbe, 2012 [33] | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 |
| Vermeulen, 2010 [30] | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 |
| Vermeulen, 2008 [36] | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 |
| Vanniyasingam, 2016 [1] | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| Graβhoff, 2013 [32] | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| Donkers, 2003 [37] | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 |

Comment: 1 = reported; 0 = unclear/not reported for each column.
*1 = the chosen design characteristics are motivated by real-world scenario (previous literature referenced, etc) OR by other simulation study scenarios, 0 = not motivated by other studies.



Ph.D. Thesis – T. Vanniyasingam; McMaster University
Health Research Methodology, Biostatistics Specialization
Contemporary Clinical Trials Communications 10 (2018) 17–28

## Appendix B. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.conctc.2018.01.002.

## References

[1] T. Vanniyasingam, C.E. Cunningham, G. Foster, L. Thabane, Simulation study to determine the impact of different design features on design efficiency in discrete choice experiments, BMJ open 6 (7) (2016) 2016–011985.
[2] M.D. Clark, D. Determann, S. Petrou, D. Moro, E.W. de Bekker-Grob, Discrete choice experiments in health economics: a review of the literature, Pharmacoeconomics 32 (9) (2014) 883–902.
[3] V.H. Decalf, A.M.J. Huion, D.F. Benoit, M.A. Denys, M. Petrovic, K. Everaert, Older People's preferences for side effects associated with antimuscarinic treatments of overactive bladder: a discrete-choice experiment, Drugs Aging (2017).
[4] J.M. Gonzalez, S. Ogale, R. Morlock, J. Posner, B. Hauber, N. Sommer, A. Grothey, Patient and physician preferences for anticancer drugs for the treatment of metastatic colorectal cancer: a discrete-choice experiment, Canc. Manag. Res. 9 (2017) 149–158.
[5] M. Feehan, M. Walsh, J. Godin, D. Sundwall, M.A. Munger, Patient preferences for healthcare delivery through community pharmacy settings in the USA: a discrete choice study, J. Clin. Pharm. Therapeut. (2017).
[6] A. Liede, C.A. Mansfield, K.A. Metcalfe, M.A. Price, C. Snyder, H.T. Lynch, S. Friedman, J. Amelio, J. Posner, S.A. Narod, et al., Preferences for breast cancer risk reduction among BRCA1/BRCA2 mutation carriers: a discrete-choice experiment, Breast Canc. Res. Treat. (2017).
[7] P. Ghijben, E. Lancsar, S. Zavarsek, Preferences for oral anticoagulants in atrial fibrillation: a best–best discrete choice experiment, Pharmacoeconomics 32 (11) (2014) 1115–1127.
[8] M. Ryan, K. Gerard, Using discrete choice experiments to value health care programmes: current practice and future research reflections, Appl. Health Econ. Health Pol. 2 (1) (2003) 55–64.
[9] M. Lagarde, D. Blaauw, A review of the application and contribution of discrete choice experiments to inform human resources policy interventions, Hum. Resour. Health 7 (1) (2009) 1.
[10] M.C. Bliemer, J.M. Rose, Experimental design influences on stated choice outputs: an empirical study in air travel choice, Transport. Res. Pol. Pract. 45 (1) (2011) 63–79.
[11] E.W. de Bekker-Grob, M. Ryan, K. Gerard, Discrete choice experiments in health economics: a review of the literature, Health Econ. 21 (2) (2012) 145–172.
[12] D. Marshall, J.F. Bridges, B. Hauber, R. Cameron, L. Donnalley, K. Fyie, F.R. Johnson, Conjoint analysis applications in health—how are studies being designed and reported? Patient Patient-Cent. Outcomes Res. 3 (4) (2010) 249–256.
[13] K.L. Mandeville, M. Lagarde, K. Hanson, The use of discrete choice experiments to inform health workforce policy: a systematic review, BMC Health Serv. Res. 14 (1) (2014) 1.
[14] L. Lourenço-Gomes, L.M.C. Pinto, J. Rebelo, Using choice experiments to value a world cultural heritage site: reflections on the experimental design, J. Appl. Econ. 16 (2) (2013) 303–332.
[15] J.J. Louviere, D. Street, L. Burgess, N. Wasi, T. Islam, A.A.J. Marley, Modeling the choices of individual decision-makers by combining efficient choice experiment designs with extra preference information, J. Choice Modell. 1 (1) (2008) 128–164.
[16] W.F. Kuhfeld, Experimental design, efficiency, coding, and choice designs, Marketing Research Methods in Sas: Experimental Design, Choice, Conjoint, and Graphical Techniques, Iowa State University: SAS Institute Inc., 2005, pp. 53–241.
[17] W.F. Kuhfeld, Marketing Research Methods in SAS. Experimental Design, Choice, Conjoint, and Graphical Techniques Cary, NC, SAS-Institute TS-722, 2005.
[18] K. Zwerina, J. Huber, W.F. Kuhfeld, A General Method for Constructing Efficient Choice Designs, Fuqua School of Business, Duke University, Durham, NC, 1996.
[19] O. Toubia, J.R. Hauser, On managerially efficient experimental designs, Market. Sci. 26 (6) (2007) 851–858.
[20] R. Kessels, P. Goos, M. Vandebroek, A comparison of criteria to design efficient choice experiments, J. Market. Res. 43 (3) (2006) 409–419.
[21] F. Carlsson, P. Martinsson, Design techniques for stated preference methods in health economics, Health Econ. 12 (4) (2003) 281–294.
[22] Z. Sandor, M. Wedel, Profile construction in experimental choice designs for mixed logit models, Market. Sci. 21 (4) (2002) 455–475.
[23] N. Arora, J. Huber, Improving parameter estimates and model prediction by aggregate customization in choice experiments, J. Consum. Res. 28 (2) (2001) 273–283.
[24] A. Burton, D.G. Altman, P. Royston, R.L. Holder, The design of simulation studies in medical statistics, Stat. Med. 25 (24) (2006) 4279–4292.
[25] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1) (1977) 159–174.
[26] D. Hensher, Reducing sign violation for vtts distributions through endogenous recognition of an Individual's attribute processing strategy, Int. J. Transp. Econ./Rivista Int. di Econ. dei Trasporti 34 (3) (2007) 333–349.
[27] R.P. Merges, Uncertainty and the standard of patentability, High Technol. Law J. 7 (1) (1992) 1–70.
[28] L.W. Sumney, R.M. Burger, Revitalizing the U.S. semiconductor industry, Issues Sci. Technol. 3 (4) (1987) 32–41.
[29] Q. Liu, Y. Tang, Construction of heterogeneous conjoint choice designs: a new approach, Market. Sci. 34 (3) (2015) 346–366.
[30] B. Vermeulen, P. Goos, M. Vandebroek, Obtaining more information from conjoint experiments by best–worst choices, Comput. Stat. Data Anal. 54 (6) (2010) 1426–1433.
[31] J. Yu, P. Goos, M. Vandebroek, Efficient conjoint choice designs in the presence of respondent heterogeneity, Market. Sci. 28 (1) (2009) 122–135.
[32] U. Graßhoff, H. Großmann, H. Holling, R. Schwabe, Optimal design for discrete choice experiments, J. Stat. Plann. Inference 143 (1) (2013) 167–175.
[33] M. Crabbe, M. Vandebroek, Improving the efficiency of individualized designs for the mixed logit choice model by including covariates, Comput. Stat. Data Anal. 56 (6) (2012) 2059–2072.
[34] M.C.J. Bliemer, J.M. Rose, Construction of experimental designs for mixed logit models allowing for correlation across choice observations, Transp. Res. Part B Methodol. 44 (6) (2010) 720–734.
[35] B. Vermeulen, P. Goos, M. Vandebroek, Rank-order choice-based conjoint experiments: efficiency and design, J. Stat. Plann. Inference 141 (8) (2011) 2519–2531.
[36] B. Vermeulen, P. Goos, M. Vandebroek, Models and optimal designs for conjoint choice experiments including a no-choice option, Int. J. Res. Market. 25 (2) (2008) 94–103.
[37] B. Donkers, P.H. Franses, P.C. Verhoef, Selective sampling for binary choice models, J. Market. Res. 40 (4) (2003) 492–497.
[38] B.J. Kanninen, Optimal design for multinomial choice experiments, J. Market. Res. 39 (2) (2002) 214–227.
[39] J.F. Bridges, A.B. Hauber, D. Marshall, A. Lloyd, L.A. Prosser, D.A. Regier, F.R. Johnson, J. Mauskopf, Conjoint analysis applications in health—a checklist: a report of the ISPOR good research practices for conjoint analysis task force, Value Health 14 (4) (2011) 403–413.
[40] F.R. Johnson, E. Lancsar, D. Marshall, V. Kilambi, A. Mühlbacher, D.A. Regier, B.W. Bresnahan, B. Kanninen, J.F. Bridges, Constructing experimental designs for discrete-choice experiments: report of the ISPOR conjoint analysis experimental design good research practices task force, Value Health 16 (1) (2013) 3–13.
[41] V. Faustin, A.A. Adégbidi, S.T. Garnett, D.O. Koudandé, V. Agbo, K.K. Zander, Peace, health or fortune?: Preferences for chicken traits in rural Benin, Ecol. Econ. 69 (9) (2010) 1848–1857.
[42] B.K. Orme, Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research, Research Publishers, 2010.

## Supplementary Files

Table 1: Screening Process Dates of studies (From inception – July 20, 2016)

| **Round 1 screening** | | |
|---|---|---|
| **Database (Search date)** | **Results** | **Search Terms** |
| **JSTOR** (From inception – July 20, 2016) | 2,604,207 | Field 1: ((discrete choice experiment*) OR ("discrete choice" ) OR (conjoint analysis) OR ("latent class")) in Full Text |
| | 842 | Field 2: (("design efficiency" OR "D-efficiency" )) in Full Text |
| | 162,971 | Field 3: ((simulation* OR "simulated design" OR "simulation study")) in Full Text |
| | 186 | **Combine Fields 1 AND Field 2 AND Field 3 in Full Text**<br>((discrete choice experiment*) OR ("discrete choice" ) OR (conjoint analysis) OR ("latent class")) AND (("design efficiency" OR "D-efficiency" )) AND<br>((simulation* OR "simulated design" OR "simulation study")) |
| **Science Direct** (From inception – July 20, 2016) | 24,287 | Field 1: (("stated preference*" OR "latent class analysis" OR "latent class" OR "conjoint analyses" OR "conjoint analysis" OR "choice experiment" OR "choice behaviour" OR "discrete choice" OR "discrete choice experiment") ) in All Fields |
| | 1,369 | Field 2: ((("simulated design" OR "simulation study" OR (simulation*)) AND ("D-optimality" OR "design optimality" OR "D-efficiency" OR "design efficiency"))) in All Fields |
| | 111 | **Combine Fields 1 AND 2:**<br>(("stated preference*" OR "latent class analysis" OR "latent class" OR "conjoint analyses" OR "conjoint analysis" OR "choice experiment" OR "choice behaviour" OR "discrete choice" OR "discrete choice experiment") )<br>AND<br>((("simulated design" OR "simulation study" OR (simulation*)) AND ("D-optimality" OR "design optimality" OR "D-efficiency" OR "design efficiency"))) |
| **Pubmed** (From inception – July 20, 2016) | 5389 | Field 1: ("stated preference*" OR "latent class analysis" OR "latent class" OR "conjoint analyses" OR "conjoint analysis" OR "choice experiment" OR "choice behaviour" OR "discrete choice" OR "discrete choice experiment") |
| | 350843 | Field 2: (("simulated design" OR "simulation study" OR (simulation*)) |
| | 173 | Field 3: ("D-optimality" OR "design optimality" OR "D-efficiency" OR "design efficiency") |
| | 2 | **Combine all 3 fields:**<br>(((((("stated preference*" OR "latent class analysis" OR "latent class" OR "conjoint analyses" OR "conjoint analysis" OR "choice experiment" OR "choice behaviour" OR "discrete choice" OR "discrete choice experiment")))) AND<br>(((("simulated design" OR "simulation study" OR (simulation*))))) AND ((("D-optimality" OR "design optimality" OR "D-efficiency" OR "design efficiency"))) |
| **OVID** (From inception – July 20, 2016) | 47917 | Field 1: ("stated preference*" OR "latent class analysis" OR "latent class" OR "conjoint analyses" OR "conjoint analysis" OR "choice experiment" OR "choice behaviour" OR "discrete choice" OR "discrete choice experiment") |
| | 1202337 | Field 2: ("simulated design" OR "simulation study" OR simulation*) |
| | 2082 | Field 3: ("D-optimality" OR "design optimality" OR "D-efficiency" OR "design efficiency") |
| | 51 | **Combine all 3 fields:**<br>(("stated preference*" OR "latent class analysis" OR "latent class" OR "conjoint analyses" OR "conjoint analysis" OR "choice experiment" OR "choice behaviour" OR "discrete choice" OR "discrete choice experiment")<br>AND<br>("simulated design" OR "simulation study" OR (simulation*)<br>AND<br>("D-optimality" OR "design optimality" OR "D-efficiency" OR "design efficiency")).af. |
| **Round 2 screening** | | |
| **Database (Search dates)** | **Results** | **Search Terms** |
| **JSTOR** (From July 20, 2016 – Sept 19, 2016) | 381 | Field 1: same as Round 1 |
| | 0 | Field 2: same as Round 1 |
| | 14 | Field 3: same as Round 1 |
| | 0 | **Combine Fields 1 AND Field 2 AND Field 3 in Full Text:** same as Round 1 |
| **Science Direct** (2016-present) | 2,287 | Field 1: same as Round 1 |
| | 103 | Field 2: same as Round 1 |
| | 2 (total 7, 5 from previous search) | **Combine Fields 1 AND 2:**same as Round 1 |
| **PubMed** (2016-present) | 5532 | Field 1: same as Round 1 |
| | 356460 | Field 2: same as Round 1 |
| | 178 | Field 3: same as Round 1 |
| | 2 (total 3, 1 from previous search) | **Combine all 3 fields:** same as Round 1 |
| **OVID** (2016 – current) | 2726 | Field 1: same as Round 1 |
| | 48326 | Field 2: same as Round 1 |
| | 82 | Field 3: same as Round 1 |
| | 0 (Total 3: 2 duplicates, 2 from last search, 1 found in from re-searching pubmed) | **Combine all 3 fields:** same as Round 1 |

Table 2: Details of the design characteristics explored by study

| **Design Characteristic** | **(First author, publication year)** | **Range** |
|---|---|---|
| Choice tasks | (Vanniyasingam, 2016)[1] | 2-20 |
| Attributes | (Vanniyasingam, 2016)[1] | 2-20 |
| Alternatives | (Vermeulen, 2010)[30] | 4,5,6 |
| | (Vanniyasingam, 2016)[1] | 2-5 |
| Attribute levels | (Vanniyasingam, 2016)[1] | 2-5 |
| | (Graβhoff, 2013)[32] | Unrestricted quantitative variable and qualitative variables |
| Number of different attribute levels between alternatives | (Graβhoff, 2013)[32] | 0-7 |
| Incorporating choice behaviour within design creation | (Donkers, 2003)[37] | For binary choice designs, % of population assumed to choose one alternative: 2.5%, 5%, 10%,15%, 150% |
| | (Crabbe, 2012)[33] | Individually adapted sequential Bayesian (IASB) design incorporating covariates; IASB not incorporating covariates |
| Priors | (Yu, 2009)[31] | • semi-Bayesian mixed logit designs (with varying heterogeneity priors),<br>• locally d-optimal mixed logit model,<br>• Bayesian multinomial logit d-optimal model (with different covariance priors),<br>• locally d-optimal design,<br>• multinomial logit model,<br>• nearly orthogonal mode<br><br><u>Heterogeneity priors:</u><br>• No heterogeneity prior specified • 0.5*[1 1 … 1]'<br>• [0 0 … 0]' • 1.5*[1 1 … 1]'<br>• [1 1 … 1]' |
| | (Vermeulen, 2010)[30] | Parameter variance priors:<br>• 0.04 • 0.5 |
| | (Bliemer, 2010)[34] | Misspecification of parameter priors |
| Methods to create design<br>- *comparing against designs* | (Vermeulen, 2010)[30] | • semi-Bayesian d-optimal best-worst choice design (main comparator);<br>• utility-neutral best-worst choice design;<br>• semi-Bayesian d-optimal choice design;<br>• utility-neutral choice design; nearly orthogonal design; random design;<br>• balanced attribute level overlap design |
| Methods to create design<br>- *in various design settings* | (Bliemer, 2010)[34] | • multinomial logit model<br>• cross-sectional mixed logit<br>• panel mixed logit<br>• orthogonal design (orthogonal within alternatives, not across alternatives) |
| | (Vermeulen, 2010)[30] | • best choice experiment<br>• partial rank-order conjoint experiment<br>• best worst experiment<br>• orthogonal design<br>All designs were created and compared within each design setting (note: orthogonal design was not also assumed as a design setting) |
| | (Vermeulen, 2008)[36] | • multinomial logit (ignores no-choice option)<br>• extended no-choice multinomial logit<br>• nested no-choice multinomial logit<br>• nested no-choice multinomial logit<br>• model robust models |
| | (Vermeulen, 2011)[35] | • d-optimal rank design<br>• d-optimal choice design<br>• balanced overlap design<br>• near-orthogonal design<br>• random design |

Open Access Research

BMJ Open

# Simulation study to determine the impact of different design features on design efficiency in discrete choice experiments

Thuva Vanniyasingam,[1] Charles E Cunningham,[2] Gary Foster,[1,3] Lehana Thabane[1,3,4,5,6]

**To cite:** Vanniyasingam T, Cunningham CE, Foster G, *et al*. Simulation study to determine the impact of different design features on design efficiency in discrete choice experiments. *BMJ Open* 2016;**6**:e011985. doi:10.1136/bmjopen-2016-011985

▸ Prepublication history for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2016-011985).

Received 23 March 2016
Revised 16 May 2016
Accepted 24 May 2016

For numbered affiliations see end of article.

**Correspondence to**
Thuva Vanniyasingam;
thuva.vanni@gmail.com

## ABSTRACT

**Objectives:** Discrete choice experiments (DCEs) are routinely used to elicit patient preferences to improve health outcomes and healthcare services. While many fractional factorial designs can be created, some are more statistically optimal than others. The objective of this simulation study was to investigate how varying the number of (1) attributes, (2) levels within attributes, (3) alternatives and (4) choice tasks per survey will improve or compromise the statistical efficiency of an experimental design.

**Design and methods:** A total of 3204 DCE designs were created to assess how relative design efficiency (d-efficiency) is influenced by varying the number of choice tasks (2–20), alternatives (2–5), attributes (2–20) and attribute levels (2–5) of a design. Choice tasks were created by randomly allocating attribute and attribute level combinations into alternatives.

**Outcome:** Relative d-efficiency was used to measure the optimality of each DCE design.

**Results:** DCE design complexity influenced statistical efficiency. Across all designs, relative d-efficiency decreased as the number of attributes and attribute levels increased. It increased for designs with more alternatives. Lastly, relative d-efficiency converges as the number of choice tasks increases, where convergence may not be at 100% statistical optimality.

**Conclusions:** Achieving 100% d-efficiency is heavily dependent on the number of attributes, attribute levels, choice tasks and alternatives. Further exploration of overlaps and block sizes are needed. This study's results are widely applicable for researchers interested in creating optimal DCE designs to elicit individual preferences on health services, programmes, policies and products.

## Strengths and limitations of this study

- The statistical efficiency of various fractional factorial designs using full profiles was explored.
- The study allows identification of optimal designs with reduced response burden for participants.
- The results of this study can be used in designing discrete choice experiments (DCEs) studies to better elicit preferences for health products and services.
- Statistical efficiency of partial profile designs was not explored.
- Optimal DCE designs require a balance between statistical efficiency and response burden.

## INTRODUCTION

Determining preferences of patients and healthcare providers is a critical approach to providing high-quality healthcare services. Discrete choice experiments (DCEs) are a relatively easy and inexpensive approach to determining the relative importance of aspects in decision-making related to health outcomes and healthcare services.[1–15] DCEs have long been applied in market research,[16–21] while health research has more recently recognised their usefulness. With increasing popularity and a wide variety of applications, few studies have investigated the effect of multiple design characteristics on the statistical efficiency of DCEs.

In practice, DCEs are presented as preference surveys where respondents are asked to choose from two or more alternatives. These alternatives are bundles of multiple attributes that describe real-world alternatives.[22] They are randomly placed within choice tasks (ie, survey questions) to create a survey where participants are asked to choose their most preferred option. Based on the alternatives chosen, the value of participant preferences on each attribute and attribute level can then be measured using the random utility theory.[22] The ratios of these utility measures are used to compare factors with different units.

For DCE designs exploring a large number of variables, where presenting all combinations of alternatives is not feasible, a fractional factorial design can be used to determine participant preferences. For example, Cunningham *et al*[15] investigated the most preferred knowledge translation approaches among individuals working in addiction agencies for women. They investigated 16 different four-level knowledge dissemination variables in a preference survey of 18 choice tasks, three alternatives per choice task, and 999 blocks. Blocks are surveys containing a different set of choice tasks (ie, presenting different combinations of alternatives), where individuals are randomly assigned to a block.[15] To create a full factorial design with 16 four-level attributes, a total of 4 294 967 296 ($4^{16}$) different hypothetical alternatives are needed. Cunningham *et al* created a design with 999 blocks of 18 choice tasks and three alternatives per choice task. In total, this was a collection of 53 946 hypothetical scenarios, <1% of all possible scenarios.

When a small fraction of all possible scenarios is used in a DCE, biased results may occur due to how evenly attributes are represented. A full-factorial design presents all possible combinations of attributes and attribute-levels to participants. Such a design achieves optimal statistical efficiency; however, it is not usually practical or feasible to implement. Fractional factorial designs are pragmatic and present only a fraction of all possible choice tasks, but statistical efficiency is compromised in the process. The goodness of a fractional factorial design is often measured by relative design efficiency (d-efficiency), a function of the variances and covariances of the parameter estimates.[23] A design is considered statistically efficient when its variance–covariance matrix is minimised.[23] Poorly designed DCEs may lead to poor data quality, potentially leading to less reliable statistical estimates or erroneous conclusions. A less efficient design may also require a larger sample size, leading to increased costs.[24 25] Investigating DCE design characteristics and their influence on statistical efficiency will aid investigators in determining appropriate DCE designs.

Previous studies have taken various directions to explore statistical efficiency, either empirically or with simulated data. These approaches (1) identified optimal designs using specific design characteristics,[26–28] (2) compared different statistical optimality criteria,[29 30] (3) explored prior estimates for Bayesian designs[31–34] and (4) compared designs with different methods to construct a choice task (such as random allocation, swapping, cycling, etc).[25 29 35–37] Detailed reports have been produced to describe the key concepts behind DCEs such as their development, design components, statistical efficiency and analysis.[38 39] However these reports did not address the effect of having more attributes or more alternatives on efficiency.

To assess previous work in this area, we conducted a literature review of DCE simulation studies. Details are reported in box 1. In our search, the type of outcome differed across studies, making it difficult to compare results and identify patterns. We focused on relative d-efficiency (or d-optimality) and also reviewed a couple of studies that reported d-error, an inverse of relative d-efficiency.[40 41] Of the studies reviewed, the various design characteristics explored by simulation studies are presented in table 1. Within each study, only two to three characteristics were explored. The number of alternatives investigated ranged from 2 to 5, attributes from 2 to 12, and attribute levels from 2 to 7. Only one study compared different components of blocks.[42] To our knowledge, no study has investigated the impact of multiple DCE characteristics with pragmatic ranges on statistical efficiency.

The primary objective of this paper is to determine how the statistical efficiency of a DCE, measured with relative d-efficiency, is influenced by various experimental design characteristics including the number of: choice tasks, alternatives, attributes and attribute levels.

**Box 1** Search strategy for reviews on applications of DCEs in health literature

A systematic search was performed using the following databases and search words. Snowball sampling was also performed in addition to the systematic search.

*Databases searched:*

- JSTOR, Science Direct, PubMed and OVID.

*Search words (where possible, given restrictions of each database)*

- dce,
- discrete choice,
- discrete-choice,
- discrete choice experiment(s),
- discrete choice conjoint experiment(s),
- discrete choice modelling/modelling,
- choice behaviour,
- choice experiment,
- conjoint analysis/es,
- conjoint measurement,
- conjoint choice experiment(s),
- latent class,
- stated preference(s),
- simulation(s),
- simulation study,
- simulated design(s),
- design efficiency,
- d-efficiency,
- design optimality,
- d-optimality,
- relative design efficiency,
- relative d-efficiency,
- relative efficiency.

## METHODS

DCEs are attribute-based approaches that rely on two assumptions: (1) products, interventions, services or policies can be represented by their attributes (or

**Table 1** Design characteristics investigated by simulation studies

| | First author, year | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Design characteristic** | **Street[28] 2002** | **Kanninen[27] 2002** | **Demirkale[42] 2013** | **Graßhoff[47] 2013** | **Louviere[24] 2008** | **Crabbe[40] 2012** | **Vermeulen[48] 2010** | **Donkers[41] 2003** | **This study** |
| Number of choice tasks | 8–1120* | 360 | | Varied to achieve optimality | 4,8,16,32* | 16 | 9 | | 2–20* |
| Number of alternatives | 2 | 2,3,5* | 2,3* | 3 | 2 | 3 | 5 | 2 | 2–5* |
| Number of attributes | 3–8* | 2,4,8* | 3–12* | 1–7* | 3–7* | 3 | 2,3* | 2 | 2–20* |
| Number of levels | 2 | 2 | 2–7* | 2 | 1,2 | 3 | | 2 | 2–5* |
| Number of blocks | | | 5 | | | | | | |
| Sample size | | | | | 38–106* | 25, 250* | 50 | | |
| Outcome type | D-efficiency | D-optimality | Number choice sets to achieve d-optimality | D-efficiency | D-efficiency | D-error | Relative d-efficiency | D-error | Relative d-efficiency |
| Comments | Only 38 designs presented. | Attribute levels described by as lower and upper bound | Evaluate different components of blocks | Locally optimal designs created. Compared binary attributes with 1 quantitative attribute, swapped alternatives within choice sets | Variation of levels is referred to as level differences | Authors compared designs with and without covariate information | Compared best-worst mixed designs with designs that were: (1) random, (2) orthogonal, (3) with minimal overlap, (4) d-optimal and (5) utility neutral d-optimal design | Designs compared with a binary attribute with an even distributed vs a skewed distribution | Characteristics were individually varied, holding others constant, to explore their impact on relative d-efficiency |

*Design characteristic has been investigated.

characteristics); and (2) an individual's preferences depend on the levels of these attributes.[14] Random allocation was used to place combinations of attributes and attribute levels into alternatives within choice tasks.

### Process of creating multiple designs

To create each design, various characteristics of DCEs were explored to investigate their impact on relative d-efficiency. The basis of each characteristic's range was determined by literature reviews and systematic reviews of applications of DCEs (table 2). The reviews covered DCE studies from 1990 to 2013, exploring areas such as economic evaluations, transportation and healthcare. The number of choice tasks per participant was most frequently 20 or less, with 16 or fewer attributes, between two and seven attribute levels, and between two and six alternatives. While the presence of blocks was reported, however, the number of blocks in each study was not.

Using the modes of design characteristics from these reviews, we simulated 3204 DCE designs. A total of 288 (18×4×4=288) designs were created to determine how relative d-efficiency varied with 2–20 attributes, 2–5 attribute levels, and 2–5 alternatives. Each of the 288 designs had 20 choice tasks. We then continued to explore designs with different numbers of choice tasks. A total of 2916 (18×18×3×3=2916) designs were created that ranged with choice tasks from 2 to 20, attributes from 2 to 20, attribute levels from 2 to 4 and alternatives from 2 to 4.

### Generating full or fractional factorial DCE designs in SAS V.9.4

The generation of full and fractional factorial designs was created using generic attributes in V.9.4 SAS software (Cary, North Carolina, USA). Four built-in SAS macros (%MktRuns, %MktEx, %MktLab and %ChoiceEff) are typically used to randomly allocate combinations of attributes and attribute levels to generate optimal designs.[43] The %MktEx macro was used to create hypothetical combinations of attributes and attribute levels in a linear arrangement. Alternatives were added with %MktLab, results were assessed and then transformed into a choice design using %ChoiceEff.[43]

### Evaluating the optimality of the DCE design

To evaluate each choice design, the goodness or efficiency of each experimental design was measured using relative d-efficiency. It ranges from 0% to 100% and is a relative measure of hypothetical orthogonal designs. A d-efficient design will have a value of 100% when it is balanced and orthogonal. Values between 0% and 100% indicate that all parameters are estimable, however, will have less precision than an optimal design. D-efficiency measures of 0 indicate that one or more parameters cannot be estimated.[43] Designs are balanced when the levels of attributes appear an equal number of times in choice tasks.[3 43] Designs are orthogonal when there is equal occurrence of each possible pair of levels across all pairs of attributes within the design.[43] Since full factorial designs present all possible combinations of attributes and attribute levels, they are always balanced and orthogonal with a 100% d-efficiency measure. Fractional factorial designs present only a portion of these combinations, creating variability in statistical efficiency.

## RESULTS

A total of 3204 simulated DCE designs were created, varying by several DCE design characteristics. Using these designs, we present the impact of each design characteristic on relative d-efficiency by the number of alternatives, attributes, attribute levels and choice tasks in a DCE, respectively.

Relative d-efficiency increases with more alternatives per choice task in a design. This was consistent across all designs with various numbers of attributes, attribute levels and choice tasks. Figure 1A–D displays this change in statistical optimality for designs with two, three, four and five alternatives ranging from 2-level to 5-level attributes, 2 to 20 attributes, and a choice set size of 20. The same effect is found on designs across all choice set sizes ranging from 2 to 20.

As the number of attributes increases, relative d-efficiency decreases, and in some cases designs were not producible. Designs with a larger number of attributes could not be created with a small number of alternatives or choice tasks. Figure 2A displays the decline in relative d-efficiency with DCEs ranging from two to five attributes across 2 to 20 choice tasks. Figure 2B–D illustrates a larger decline in relative d-efficiency as attribute size increases from 6 to 10, 11 to 15 and 16 to 20, respectively. Designs with choice tasks less than 11 were not possible in these examples.

Similarly, from comparing figure 2B with figure 3, as the number of attribute levels increase, relative d-efficiency decreases across all designs with varying numbers of attributes, choice tasks and alternatives. DCEs with binary attributes (figure 2B) consistently performed well with all relative d-efficiencies above 80% except for designs with 18 or more attributes.

As the number of choice tasks in a design increases, d-efficiency increases and may plateau, where this plateau may not reach 100% statistical efficiency. This was observed across all attributes and attribute levels. Relative d-efficiency peaked at designs with a specific number of choice tasks, particularly when the number of alternatives was equal to or a multiple of the number of attribute levels and the number of choice tasks. This looping pattern of peaks begins only at large choice set sizes for designs with a large number of attributes. For example, among designs with two alternatives and two-level attributes, peaks were observed for designs with choice set sizes as small as 2 (figure 2A,B). For designs with three alternatives and three-level attributes, this looping pattern appeared at choice set sizes of 3, 9, 12, 15 and 18, depending on how much larger or smaller the number of attributes was.

Table 2 Summary of items reported by reviews of DCEs

| First author | Ryan[13] | Lagarde[49] | Marshall[1] | Bliemer[44] | de Bekker-Grob[3] | Mandeville[2] | de Bekker-Grob[25] | Clark[50] |
|---|---|---|---|---|---|---|---|---|
| **Description of reviews** | | | | | | | | |
| Year reported | 2003 | 2009 | 2010 | 2011 | 2012 | 2014 | 2015 | 2014 |
| Years covered | 1990–2000 | No time limit | 2005–2008 | 2000–2009 | 2001–2008 | 1998–2013 | 2012 | 2009–2012 |
| Literature review (LR) or systematic review (SR) | LR | LR | SR | LR | SR | SR | LR | SR |
| Specialities, areas covered in review | Healthcare, economic evaluations, other (eg, insurance plans) | Health workers | Disease-specific primary health studies | Tier 1 transportation journals | Health economics, QALY | Labour market preferences of health workers/ human resources for health | Sample size calculations for healthcare-related DCE studies | Health-related DCEs |
| Total number of studies assessed | 34 | 10 | 79 | 61 | 114 | 27 | 69 | 179 |
| **Items reported** | | | | | | | | |
| Number of choice tasks given to each participant | <8, 9–16, >16, not reported (mode=9–16) | Only reported mode 16 | 2–35, not reported (mode=7) | 1–20, not reported (mode=8,9) (total across all blocks: 3–46) | <8, 9–16, >16, not reported (mode ≤8) | <10–20 (mode=16–20) | ≤8 to ≥16, not reported (mode=9–16) | <9 to >16 (mode=9–16) |
| Number of attributes | 2–24 (mode=6) | 5–7 | 3–16 (mode=6, 70% between 3 and 7) | 2–30 (mode=5) | 2 to >10 | 5–8 | 2–9, >9 (mode=6) | 2–>10 (mode=6) |
| Number of levels within attributes | | 2–6 | 2,3 | 2–7 | | 2–4 (mode=2) | | |
| Number of alternatives | 2, >2 | 2 | | 2–6 | 2 | 2–4 | | |
| Number of blocks | | | | Blocking reported, number of blocks not reported | | Blocking reported, number of blocks not reported | | |
| Reported DCEs using Bayesian methods | | | | | | Yes | | Yes |
| Design type: 1=full-factorial 2=fractional factorial 3=not reported | 1, 2, 3 | 2 | | 1, 2, 3 | 1, 2, 3 | 2 | | 1, 2, 3 |
| Sample size | | | 13–1258 | 20–5829 | | 102–3727 | <100 to >1000 | |
| Overlaps in alternatives | | | | | | Yes | | |
| Number of simulation studies | | | | | | | | |
| Response rates | <30–100% | | | | | 16.8–100% | | |
| Comments | | | | | Comparison with old SR (an updated SR) | A systematic update of Lagarde *et al*'s[49] study | Sample size paper | This is a systematic update of de Bekker-Grob *et al*'s[3] study |

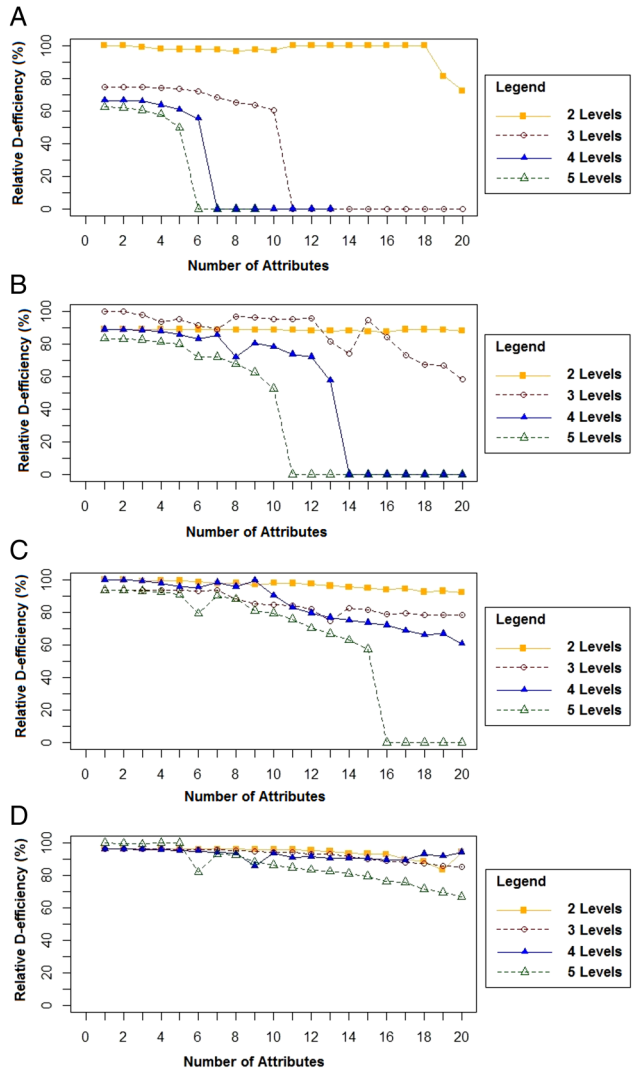



**Figure 1** (A) Relative d-efficiencies (%) of designs with two alternatives across 2–20 attributes, 2–5 attribute levels and 20 choice sets each. (B) Relative d-efficiencies (%) of designs with three alternatives across 2–20 attributes, 2–5 attribute levels and 20 choice sets each. (C) Relative d-efficiencies (%) of designs with four alternatives across 2–20 attributes, 2–5 attribute levels and 20 choice sets each. (D) Relative d-efficiencies (%) of designs with five alternatives across 2–20 attributes, 2–5 attribute levels and 20 choice sets each.

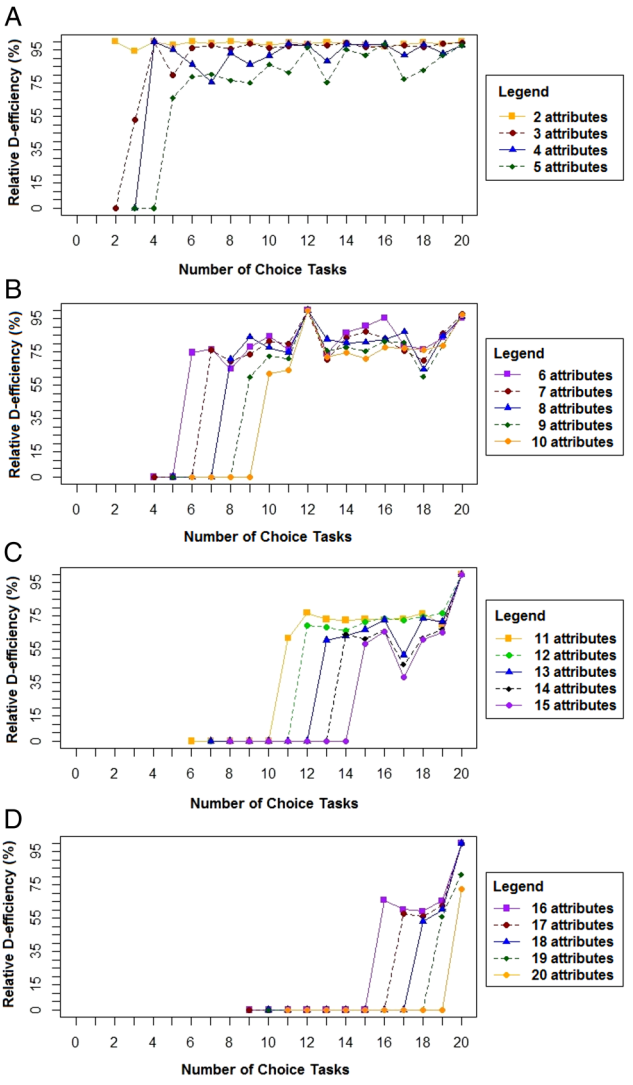



**Figure 2** (A) The effect of 2–5 attributes on relative d-efficiency (%) across different choice tasks for designs with two alternatives and two-level attributes. (B) The effect of 6–10 attributes on relative d-efficiency (%) across different choice tasks for designs with two alternatives and two-level attributes. (C) The effect of 11–15 attributes on relative d-efficiency (%) across different choice tasks for designs with two alternatives and two-level attributes. (D) The effect of 16–20 attributes on relative d-efficiency (%) across different choice tasks for designs with two alternatives and two-level attributes.

## DISCUSSION

A total of 3204 DCE designs were evaluated to determine the impact of the different numbers of alternatives, attributes, attribute levels, and choice tasks on the relative d-efficiency of a design. Designs were created by varying one characteristic while holding others constant. Relative d-efficiency increased with more alternatives per choice task in a design, but decreased as the number of attributes and attribute levels increased. When the number of choice tasks in a design increased, d-efficiency would either increase or plateau to a maximum value, where this plateau may not reach 100% statistical efficiency. A pattern of peaks in 100% relative d-efficiency occurred for many designs where the number of alternatives was equal to, or a multiple of, the number of choice tasks and attribute levels.

The results of this simulation study are in agreement with other methodological studies. Sandor *et al*[35] showed that DCE designs with a larger number of alternatives (three or four) performed more optimally using Monte Carlo simulations, relabelling, swapping and cycling techniques. Kanninen *et al*[27] emphasise the use of binary attributes and suggest optimal designs, regardless of the number of attributes. We observed a pattern where many designs achieved statistical optimality, and when the number of choice tasks is a multiple of the number of

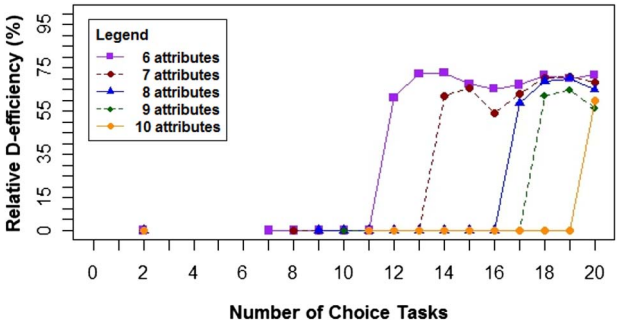



**Figure 3** The effect of 6–10 attributes on relative d-efficiency (%) across different choice tasks for designs with two alternatives and three-level attributes.

alternatives and attribute levels, relative d-efficiency will peak to 100%. Johnson *et al*[38] similarly discuss how designs require the total number of alternatives to be divisible by the number of attribute levels to achieve balance, a critical component of relative d-efficiency.

While fewer attributes and attribute levels were found to yield higher relative d-efficiency values, there is a lot of variability among applications of DCE designs (table 2). In our assessment of literature and systematic reviews from 2003 to 2015, some DCEs evaluated up to 30 attributes or 7 attribute levels.[44] De Bekker-Grob *et al*[3] observed DCEs within health economics literature between two time periods: 1990–2000 and 2001–2008. The total number of applications of DCEs increased from 34 to 114, while the proportions among design characteristics were similar. A majority of designs used 4–6 attributes (55% in 1990–2000, 70% in 2001–2008). In the 1990s, 53% used 9–16 choice tasks per design. This reduced to 38% in the 2000s with more reporting only eight or less choice tasks per design. While d-efficiency is advocated as a criterion for evaluating DCE designs,[45] it was not commonly reported in the studies (0% in 1990–2000, 12% in 2001–2008). Other methods used to achieve orthogonality were single profiles (with binary choices), random pairing, pairing with constant comparators, or a fold-over design. Following this study, de Bekker-Grob performed another review in 2012 of 69 healthcare-related DCEs, where 68% used 9–16 choice tasks and only 20% used 8 or less.[25] Marshall *et al*'s review reported many DCEs created designs with six or fewer attributes (47/79), 7–15 choice tasks (54/79), with two-level (48/79) or three-level (42/79) attributes. Among these variations, de Bekker-Grob *et al*[3] mention 37% of studies (47/114) did not report sufficient detail of how choice sets were created, which leads us to question if there is a lack of guidance in the creation and reporting of DCE designs.

This simulation study explores the statistical efficiency of a variety of both pragmatic and extreme designs. The diversity in our investigation allows for an easy assessment of patterns in statistical efficiency that is affected by specific characteristics of a DCE. We found that designs with binary attributes or a smaller number of attributes had better relative d-efficiency measures, which will also reduce cognitive burden, improve choice consistency and overall improve respondent efficiency. We describe the impact of balance and orthogonality on d-efficiency by the looping pattern observed as the number of choice tasks increase. We also link our findings with what has been investigated among other simulation studies and applied within DCEs. This study's results complement the existing information on DCE in describing the role each design characteristic has on statistical efficiency.

There are some key limitations to our study that are worth discussing. Multiple characteristics of a DCE design were explored, however, further attention is needed to assess all influences on relative d-efficiency. First, the number of overlaps, where the same attribute level is allowed to repeat in more than one alternative in a choice task, was not investigated. The presence of overlaps helps participants by reducing the number of comparisons they have to make. In SAS, the statistical software we used in creating our DCE designs, we were only able to specify whether or not overlaps were allowed. We were not able to specify the number of overlaps within a choice task or design so we did not include it in our analysis. Second, sample size was not explored. A DCE's statistical efficiency is directly influenced by the asymptotic variance–covariance matrix, which also affects the precision of a model's parameter estimates, and thus has a direct influence on the minimum sample size required.[25] Sample size calculations for DCEs need several components including the preferred significance level (α), statistical power level (1-β), statistical model to be used in the DCE analysis, initial belief about the parameter values and the DCE design.[25] Since the aim of this study was to identify statistically optimal DCE designs, we did not explore the impact of relative d-efficiency on sample size. Third, attributes with different levels (ie, asymmetric attributes or mixed-attribute designs) were not explored to compare with Burgess *et al*'s[26] findings. Best–worst DCEs were also not investigated. Last, we did not assess how d-efficiency may change when specifying a partial profile design to present only a portion of attributes within each alternative.

Several approaches can be made to further investigate DCE designs and relative d-efficiency. First, while systematic reviews exist on what designs are used and reported, none provide a review of simulation studies investigating statistical efficiency. Second, comparisons of optimal designs determined by different software and different approaches are needed to ensure there is agreement on statistically optimal designs. For example, the popular Sawtooth Software could be used to validate the relative d-efficiency measures of our designs. Third, further exploring the trade-off between statistical and informant (or respondent) efficiency will help tailor simulation studies to assess more pragmatic designs.[46] Informant efficiency is a measurement error caused by participants' inattentiveness when choosing alternatives, or by other

unobserved, contextual influences.[38] Using a statistically efficient design may result in a complex DCE, increasing the cognitive burden for respondents and reducing the validity of results. Simplifying designs can improve the consistency of participants' choices which will help yield lower error variance, lower choice variability, lower choice uncertainty and lower variance heterogeneity.[24] For investigators, it is best to consider balancing both statistical and informant efficiency when designing DCEs. Given our results, one approach to reduce design complexity we propose is to reduce the number of attributes and attribute levels, where possible, to identify an efficient and less complex design. Fifth, there is limited discussion of blocked DCEs among the simulation studies and reviews we explored. One study explored three different experimental designs (orthogonal with random allocation, orthogonal with blocking, and an efficient design), and found that blocking should be included in DCEs to improve the design.[36] Other studies either mentioned that blocks were used with no additional details[2 44] or only used one type of block size.[42] In SAS, a design must first be created before it can be sectioned into blocks. From our investigation, varying the number of blocks, therefore, had no impact on relative d-efficiency since designs were sectioned into different blocks only after relative d-efficiency was measured. More information can be provided from the authors upon request. A more meaningful investigation is to explore variations in block size (ie, the number of choice tasks within a block). This will change the number of total choice tasks required and impact the relative d-efficiency of a DCE. Last, investigating other real-world factors that drive DCE designs are critical in ensuring DCEs achieve optimal statistical and respondent efficiency.

## Conclusion

From the various designs evaluated, DCEs with a large number of alternatives and a small number of attributes and attribute levels performed best. Designs with binary attributes, in particular, had better statistical efficiency in comparison with other designs with various design characteristics. This study demonstrates that a fractional factorial design may achieve 100% statistical efficiency when the number of choice tasks is a multiple of the number of alternatives and attribute levels, regardless of the number of attributes. Further research needs to include investigation of the impact of overlaps, mixed attribute designs, best-worst DCEs and varying block sizes. These results are widely applicable in designing studies for determining individual preferences on health services, programmes and products. Clinicians can use this information to elicit participant preferences of therapies and treatments, while policymakers can identify what factors are important in decision-making.

**Author affiliations**
[1]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada
[2]Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, Ontario, Canada
[3]Biostatistics Unit, Father Sean O'Sullivan Research Centre, St. Joseph's Healthcare, Hamilton, Ontario, Canada
[4]Departments of Paediatrics and Anaesthesia, McMaster University, Hamilton, Ontario, Canada
[5]Centre for Evaluation of Medicine, St. Joseph's Healthcare, Hamilton, Ontario, Canada
[6]Population Health Research Institute, Hamilton Health Sciences, Hamilton, Ontario, Canada

**Acknowledgements** Warren Kuhfeld from the SAS Institute Inc. provided programming guidance for DCE design creation.

**Contributors** All authors provided intellectual content for the manuscript and approved the final draft. TV contributed to the conception and design of the study; performed the statistical analyses and drafted the manuscript; approved the final manuscript; and agrees to be accountable for all aspects of the work in relation to accuracy or integrity. LT contributed to the conception and design of the study; provided statistical and methodological support in interpreting results and drafting the manuscript; approved the final manuscript; and agrees to be accountable for all aspects of the work in relation to accuracy or integrity. CEC contributed to the conception and design of the study; critically assessed the manuscript for important intellectual content; approved the final manuscript; and agrees to be accountable for all aspects of the work in relation to accuracy or integrity. GF contributed to the interpretation of results; critically assessed the manuscript for important intellectual content; approved the final manuscript; and agrees to be accountable for all aspects of the work in relation to accuracy or integrity.

**Funding** This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests** All authors have completed the ICMJE uniform disclosure form at http://www.icmje.org/coi_disclosure.pdf. CEC's participation was supported by the Jack Laidlaw Chair in Patient-Centered Health Care.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** As this is a simulation study, complete results are available by emailing TV at thuva.vanni@gmail.com.



## REFERENCES

1. Marshall D, Bridges JF, Hauber B, *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.
2. Mandeville KL, Lagarde M, Hanson K. The use of discrete choice experiments to inform health workforce policy: a systematic review. *BMC Health Serv Res* 2014;14:367.
3. 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.
4. Ryan M, Scott DA, Reeves C, *et al*. Eliciting public preferences for healthcare: a systematic review of techniques. *Health Technol Assess* 2001;5:1–186.
5. Spinks J, Chaboyer W, Bucknall T, *et al*. Patient and nurse preferences for nurse handover-using preferences to inform policy: a discrete choice experiment protocol. *BMJ Open* 2015;5:e008941.
6. Baji P, Gulácsi L, Lovász BD, *et al*. Treatment preferences of originator versus biosimilar drugs in Crohn's disease; discrete choice experiment among gastroenterologists. *Scand J Gastroenterol* 2016;51:22–7.
7. Veldwijk J, Lambooij MS, van Til JA, *et al*. Words or graphics to present a Discrete Choice Experiment: does it matter? *Patient Educ Couns* 2015;98:1376–84.

8. McCaffrey N, Gill L, Kaambwa B, *et al*. Important features of home-based support services for older Australians and their informal carers. *Health Soc Care Community* 2015;23:654–64.
9. Veldwijk J, van der Heide I, Rademakers J, *et al*. Preferences for vaccination: does health literacy make a difference? *Med Decis Making* 2015;35:948–58.
10. Böttger B, Thate-Waschke IM, Bauersachs R, *et al*. Preferences for anticoagulation therapy in atrial fibrillation: the patients' view. *J Thromb Thrombolysis* 2015;40:406–15.
11. Adams J, Bateman B, Becker F, *et al*. Effectiveness and acceptability of parental financial incentives and quasi-mandatory schemes for increasing uptake of vaccinations in preschool children: systematic review, qualitative study and discrete choice experiment. *Health Technol Assess* 2015;19:1–176.
12. Brett Hauber A, Nguyen H, Posner J, *et al*. A discrete-choice experiment to quantify patient preferences for frequency of glucagon-like peptide-1 receptor agonist injections in the treatment of type 2 diabetes. *Curr Med Res Opin* 2015;7:1–32.
13. 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.
14. Ryan M. Discrete choice experiments in health care: NICE should consider using them for patient centred evaluations of technologies. *BMJ* 2004;328:360–1.
15. Cunningham CE, Henderson J, Niccols A, *et al*. Preferences for evidence-based practice dissemination in addiction agencies serving women: a discrete-choice conjoint experiment. *Addiction* 2012;107:1512–24.
16. Lockshin L, Jarvis W, d'Hauteville F, *et al*. Using simulations from discrete choice experiments to measure consumer sensitivity to brand, region, price, and awards in wine choice. *Food Qual Preference* 2006;17:166–78.
17. Louviere JJ, Woodworth G. Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data. *J Mark Res* 1983;20:350–67.
18. Louviere JJ, Hensher DA. Using discrete choice models with experimental design data to forecast consumer demand for a unique cultural event. *J Consum Res* 1983:348–61.
19. Haider W, Ewing GO. A model of tourist choices of hypothetical Caribbean destinations. *Leisure Sci* 1990;12:33–47.
20. Moore WL. Levels of aggregation in conjoint analysis: an empirical comparison. *J Mark Res* 1980:516–23.
21. Darmon RY. Setting sales quotas with conjoint analysis. *J Mark Res* 1979;16:133–40.
22. Louviere JJ, Flynn TN, Carson RT. Discrete choice experiments are not conjoint analysis. *J Choice Model* 2010;3:57–72.
23. Kuhfeld WF, Tobias RD, Garratt M. Efficient Experimental Design with Marketing Research Applications. *J Marketing Res* 1994;31:545–57.
24. Louviere JJ, Islam T, Wasi N, *et al*. Designing discrete choice experiments: do optimal designs come at a price? *J Consumer Res* 2008;35:360–75.
25. de Bekker-Grob EW, Donkers B, Jonker MF, *et al*. Sample size requirements for discrete-choice experiments in healthcare: a practical guide. *Patient* 2015;8:373–84. .
26. Burgess L, Street DJ. Optimal designs for choice experiments with asymmetric attributes. *J Stat Plann Inference* 2005;134:288–301.
27. Kanninen BJ. Optimal design for multinomial choice experiments. *J Marketing Res* 2002;39:214–27.
28. Street DJ, Burgess L. Optimal and near-optimal pairs for the estimation of effects in 2-level choice experiments. *J Stat Plann Inference* 2004;118:185–99.
29. Rose JM, Scarpa R. Designs efficiency for non-market valuation with choice modelling: how to measure it, what to report and why. *Aus J Agricul Res Econ* 2007;52:253–82.
30. Kessels R, Goos P, Vandebroek M. A comparison of criteria to design efficient choice experiments. *J Marketing Res* 2006;43:409–19.
31. Sandor Z, Wedel M. Designing conjoint choice experiments using managers' prior beliefs. *J Marketing Res* 2001;38:430–44.
32. Ferrini S, Scarpa R. Designs with a priori information for nonmarket valuation with choice experiments: a Monte Carlo study. *J Environ Econ Manag* 2007;53:342–63.
33. Arora N, Huber J. Improving parameter estimates and model prediction by aggregate customization in choice experiments. *J Consum Res* 2001;28:273–83.
34. Sándor Z, Wedel M. Heterogeneous conjoint choice designs. *J Marketing Res* 2005;42:210–18.
35. Sándor Z, Wedel M. Profile construction in experimental choice designs for mixed logit models. *Mark Sci* 2002;21:455–75.
36. Hess S, Smith C, Falzarano S, *et al*. Measuring the effects of different experimental designs and survey administration methods using an Atlanta managed lanes stated preference survey. *Transportation Res Rec* 2008;2049:144–52.
37. Huber J, Zwerina K. The importance of utility balance in efficient choice designs. *J Marketing Res* 1996;33:307–17.
38. Johnson F, 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.
39. Li W, Nachtsheim CJ, Wang K, *et al*. Conjoint analysis and discrete choice experiments for quality improvement. *J Qual Technol* 2013;45:74.
40. Crabbe M, Vandebroek M. Improving the efficiency of individualized designs for the mixed logit choice model by including covariates. *Comput Stat Data Anal* 2012;56:2059–72.
41. Donkers B, Franses PH, Verhoef PC. Selective sampling for binary choice models. *J Mark Res* 2003;40:492–7.
42. Demirkale F, Donovan D, Street DJ. Constructing D-optimal symmetric stated preference discrete choice experiments. *J Stat Plann Inference* 2013;143:1380–91.
43. Kuhfeld WF. Experimental design, efficiency, coding, and choice designs. Marketing Research methods in sas: Experimental design, choice, conjoint, and graphical techniques. Iowa State University: SAS Institute Inc 2005:53–241.
44. Bliemer MC, Rose JM. Experimental design influences on stated choice outputs: an empirical study in air travel choice. *Transportation Res Part A Policy Pract* 2011;45:63–79.
45. Zwerina K, Huber J, Kuhfeld WF. *A general method for constructing efficient choice designs*. Durham, NC: Fuqua School of Business, Duke University, 1996.
46. Patterson M, Chrzan K. Partial profile discrete choice: what's the optimal number of attributes. *The Sawtooth Software Conference: 2003*. Sequim, WA; 2003:173–85.
47. Graßhoff U, Großmann H, Holling H, *et al*. Optimal design for discrete choice experiments. *J Stat Plann Inference* 2013;143:167–75.
48. Vermeulen B, Goos P, Vandebroek M. Obtaining more information from conjoint experiments by best–worst choices. *Comput Stat Data Anal* 2010;54:1426–33.
49. Lagarde M, Blaauw D. A review of the application and contribution of discrete choice experiments to inform human resources policy interventions. *Hum Resour Health* 2009;7:62.
50. Clark MD, Determann D, Petrou S, *et al*. Discrete choice experiments in health economics: a review of the literature. *Pharmacoeconomics* 2014;32:883–902.

# CHAPTER 4

## Exploring the sensitivity of modelling repeated measures on the ranking of attributes' relative importance measures in discrete choice experiments

## Author Information

Thuva Vanniyasingam[1-3], Charles E Cunningham[4], Gary Foster[1,2], Amy Shi[5], Heather Rimas[4], Joanna Henderson[6], Alison Niccols[7], Maureen Dobbins[8], Wendy Sword[9], Yvonne Chen[4], Stephanie Mielko[4], Karen Milligan[10], Ellen Louise Lipman[4], Louis A. Schmidt[11], Lehana Thabane[1-3,12,13]

[1]Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, Ontario, Canada

[2]Biostatistics Unit, St Joseph's Healthcare, Hamilton, Ontario, Canada

3Departments of Paediatrics and Anaesthesia, McMaster University, Hamilton, Ontario, Canada

[4] (7)Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, ON, Canada

[5]Advanced Analytics Division, SAS Institute Inc., Cary, North Carolina, USA

[6]Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada; Centre for Addiction and Mental Health, Toronto, Ontario, Canada

[7]Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, Ontario, Canada

[8]National Collaborating Centre for Methods and Tools, McMaster University, Hamilton, Ontario, Canada

[9]School of Nursing, University of Ottawa, Ottawa, Ontario, Canada

[10]Integra, Psychology and Research, 25 Imperial Street, Toronto, Ontario, MP5 1B9, Canada

[11]Department of Psychology, Neuroscience and Behaviour, McMaster University, Hamilton, Ontario, Canada

[12]Population Health Research Institute, Hamilton Health Sciences, Hamilton, Ontario, Canada

[13]Centre for Evaluation of Medicine, St Joseph's Healthcare, Hamilton, Ontario, Canada

## Corresponding author:

Thuva Vanniyasingam; Department of Health Research Methods, Evidence and Impact, HSC 2C McMaster University, Hamilton, ON, L8S 4L8; thuva.vanni@gmail.com

**ABSTRACT**

**Background:** Attributes used in discrete choice experiments (DCEs) vary in scale, making it difficult to compare utilities between them. Relative importance scores offer a common scale across attributes, allowing for such comparisons of participants' preferences between attributes. These scores can then be ranked to identify which attributes of a product or service is most preferred over others. As participants respond to multiple choice tasks in a DCE, it is unclear how these repeated measures impact the ranking of attributes' relative importance measures.

**Objective:** We explored the robustness of the ranking of attributes' relative importance measures using four models that adjusted for repeated measures in different ways. We also explored the robustness of the ranking of levels within attributes across models.

**Methods:** Data was obtained from a previous DCE of 1371 service providers or administrators from 333 Canadian addiction agencies for women. We empirically analyzed the DCE data using four Bayesian models - namely, fixed effects multinomial logit (MNL), fixed effects multinomial probit (MNP), random effects MNL and random effects MNP models. The scale of relative importance scores is from 0% (least important) to 100% (most important). They were derived for each attribute based on the the mean part-worth utility estimates from each regression model.

**Results:** Many participants were service providers (74%), addiction professionals (59%), and had more than five years in their position (47%). The top five preferred attributes are client impact (14.74%), implementation complexity (12.83%), presenter's background (10.23%), work

compatibility (10.12%), and collaborative selection process (8.36%). This ranking was similar across all models. The rankings of levels within attributes were consistent across models.

**Conclusions:** Relative importance measures allow us to make comparisons between attributes and rank them in order from high to low preference. This empirical comparison found that the rankings of attributes and attribute levels were relatively robust across different models adjusting for repeated measures. This consistency reassures investigators about the robustness of their findings and strengthens their conclusions. Further research includes exploring the robustness of relative importance rankings for different DCE designs and different levels of clustering within the data.

**Key Words**

Addiction; knowledge translation; discrete choice experiment; multinomial logit; multinomial probit

## 1. INTRODUCTION

Evidence-based medicine has become the forefront of clinical practice, however, understanding values and preferences is needed to improve clinical decision making[1-3]. Health research is expanding its use of discrete choice experiments (DCEs) to elicit these preferences with applications in health economics[4, 5], health policy[6, 7], health care services[8, 9], and drug therapy[10, 11]. Several systematic reviews have also emerged to capture applications of DCEs within health research[12-18].

DCEs are a tool used to elicit and quantify participant preferences. Applying Lancaster's consumer theory, that goods and services can be described using their properties or characteristics[19], one can determine which characteristics participants are most interested in. Different combinations of these characteristics are created and randomly placed into alternatives within a survey question, where each question has a minimum of two alternatives to choose from. Participants are asked to choose their most favourable alternative based on the characteristics described for each question in the preference survey. In the realm of DCEs, these characteristics are referred to as attributes, survey questions are choice tasks, and alternatives are the options presented within each choice task. From all of the alternatives chosen at the end of the experiment, random utility theory can be applied to determine the maximum utility of each attribute level. These are quantified measures of the mean value participants have on each attribute level. To determine individuals' preferences of attributes relative to other attributes, we can manipulate their utility measures and rank them from the most preferred attribute to the least preferred [20].

To be able to appropriately compare between attributes, we need to ensure the attributes' utility measures are converted on the same scale. For example, two quantitative attributes such as time and cost are measured using hours and dollars, respectively. Utility estimates will not have the same meaning for each attribute. Similarly, we cannot compare these utility estimates with those estimated for qualitative or nominal variables (such as type of caregiver). To compare *between* attributes, we can transform their mean part-worth utility estimates, derived from regression analysis, into relative importance scores that range from 0% (least preferred) to 100% (most preferred) [21]. These scores are easily derived, provide a common scale across all attributes, and are intuitive to understand when communicating findings. Investigators can then use these relative importance scores to rank attributes and determine what is more preferred than others. To compare *within* attributes, the levels will share the same scale. We can therefore compare and rank levels within attributes using their mean part-worth utility estimates.

To begin the analysis of DCEs, we must first determine an appropriate regression model to derive these mean part-worth utility estimates. While DCEs are able to quantify individuals' preferences, correlated responses may arise when each participant is asked to answer more than one choice task. For feasibility, each participant often completes an entire survey, making choices on several choice tasks. As more than one choice task is completed by each individual, their responses are likely to be similar, potentially biasing the utility measures. Regression models that do not appropriately capture within- and between- participant variability are limited in their ability to adequately adjust for heterogeneity[22]. This can, in turn, potentially influence the final ranking of the relative importance participants place on attributes.

Several systematic reviews of applied DCEs in health research identify a diverse set of designs and statistical analysis methods[12-18, 23]. These details are presented in Table 1 at the end of this document, where all tables and figures are located. The analysis approaches varied from weighted least squares method to fixed effects logit/probit models to random effects or mixed effects logit/probit models. De Bekker-Grob and colleagues completed a review of two decades of DCEs from 1990 to 2000 and 2001 to 2008, showing that there is no clear increase in trend towards a specific statistical analysis approach for DCEs[15]. In comparison to the 1990-2000 decade, a larger volume of studies in 2001 to 2008 used a random effects probit or a multinomial logit model, however an increase was also found across all models. In a review of DCEs published between 2001 and 2008, Clark and colleagues [23] identified an increased use (18% to 44%) of multinomial models in comparison to studies conducted between 1990 to 2000[23]. A significant change from the 1990-2000 decade was the increased use (3% to 21%) of mixed effects logit or random parameter logit (RPL) designs. This indicates that more studies are incorporating heterogeneity in analysing DCE data[23]. In comparing these findings, however, to two reviews of DCEs during the span of two decades - 2000 to 2016[16, 18] - we continue to see both fixed and random or mixed effects models being used.

With computer software improving over the years and providing the ability to perform more complex analyses, we wonder why many studies from our review did not use random or mixed effects models to account for within- and between- clustering effects. In a fixed effects model, an assumption is made that the true effects of predictors (in both magnitude and direction) is the same value across the entire sample (that is, fixed across individuals). This indicates that the observed differences among individual results are due to random chance, meaning there is no statistical

heterogeneity. A random-effects model is used to analyze hierarchical data. In this case, where each participant made choices from several choice tasks, there may be similarities across their selections, creating clusters in the data. A random effects model assumes that the estimated effects within and between individuals are heterogeneous and follow some distribution. A traditional fixed effects model that does not recognize the multilevel structure of a dataset can lead to misleading statistical inferences. This can potentially lead to inaccurate estimates of the standard errors of regression coefficients, which then ends with incorrect conclusions of statistical significance for the regression coefficients[24]. Our assessment of the systematic reviews led us to question how utility measures of mean preferences on attribute levels differ by the statistical model used.

In Hensher and Greene's discussion of mixed models in DCEs, they recommend that a multinomial logit model always be the first step for an empirical investigation[25]. It assists with the major details of the modelling process and aids in ensuring data are clean and sensible results can be derived[25]. Also, if preference heterogeneity is present in the data, it will influence the marginal rates of substitution between attributes and lead to IIA violations. It is advised to incorporate heterogeneity in the regression model to determine accurate choice model predictions. This type of mixed logit regression model is often referred to as a random parameter logit (RPL), mixed multinomial logit, kernel logit, hybrid logit and error components logit (ECL) [25, 26] – several of which were used to describe models used from our literature review.

The aim of this study was to determine whether the final ranking of attributes and attribute levels differed depending on whether or not a model adjusted for repeated measures. Our first objective was to compare the ranking of attributes and attribute levels estimated by a 1) random

effects multinomial logit (MNL) model, 2) random effects multinomial probit (MNP) model, 3) fixed effects MNL model and a 4) fixed effects MNP model. Secondly, we wanted to identify a final ranking for attributes and attribute levels investigated by Cunningham and colleagues to determine knowledge translation variables important to service providers and administrators working at addiction agencies for women[27].

## 2. METHODS

### 2.1 Summary of study

Cunningham and colleagues conducted a DCE to explore which knowledge translation attributes were most preferred by service providers and administrators of addiction agencies for women[27]. A multidisciplinary team of experts identified 16 four-level attributes: impact on clients, quality of evidence (evidence quality), compatibility, implementation complexity, number of days for training (time cost), administrative support, source of endorsement, collaborative selection process, presenter's background, supplementary information, internet options, focus on knowledge versus skills, individuals versus group format, active versus passive-learning, and number of implementation follow-ups. All attributes and corresponding levels are described in Table 2. Sawtooth Software's SSI Web (Version 6.8) was used to produce a balanced, fractional factorial, partial profile design with blocks and overlaps. Each of the 1371 service providers and administrators from 333 addiction agencies were randomly assigned to one of 999 different surveys comprised of 20 questions (or choice tasks), two of which were fixed across all surveys to examine internal validity. Each choice task presented three professional development alternatives. Each alternative included one level from three of the 16 four-level attributes. Participants were asked to select which of the three themes in a choice task was most preferred. The Internet surveys were completed by 1379 (60%) of the 2305 individuals who received the survey link [27].

## 2.2 General Models

A total of four utility models were created using Statistical Analysis Software (SAS) Version 9.4 (SAS Institute Inc., Cary, NC, USA) to explore the mean part-worth values of the 16 dissemination attributes. All models were created using a Bayesian approach and specified in SAS with a procedure called PROC BCHOICE[28, 29]. Since the choice tasks in our empirical dataset were presented using 3 alternatives, we needed MNL and MNP models for our regression analyses. These models are used for categorical outcomes of two or more levels. A reference level for each of the four-level attributes was assigned a part-worth of zero and the remaining levels were set as contrasts with respect to zero. A posterior mean and highest posterior density (HPD) intervals (which are 95% credibility intervals) was estimated for each factor, representing the average part-worth utility measure of each attribute level in comparison to the reference category. These part-worth utility estimates provide insight into the extent to which participants prefer each level of an attribute[21]. A positive value indicates that an attribute level is preferred over the reference level, while a negative value implies the preference is higher for the reference level. The larger the part-worth utility value, the higher the preference level. Since the levels within attributes shared the same utility scale, we used their mean utility estimates to rank them from highest to lowest preference within each attribute. Factors were considered to have an association with the outcome if the 95% HPD intervals did not include zero.

## 2.3 Fixed effects multinomial logit (MNL) model

Fixed effects MNL models in DCEs are used to estimate the log odds that an individual will select a specific alternative, given a set of attributes and attribute levels. The outcome is the chosen alternative from each choice task and the predictors are the attributes and attribute levels.

MNL models are based on the independence of irrelevant alternatives (IIA) assumption which assumes there is no correlation of alternatives across choices. This implies that MNL is appropriate when the error terms are identical and independently distributed, with a constant variance and no correlation across alternatives [30]. The deviance information criterion (DIC) is a model assessment tool that was used to compare the model fit of the fixed effects MNL model and the random-effects only MNL model. The smaller the DIC, the better fit a model is to the dataset[31].

**2.4 Fixed effects multinomial probit (MNP) models**

In contrast to MNL models, MNP models relax the IIA assumption. This is useful when the error components are not identical or independently associated[30, 32]. Instead, they are assumed to follow a multivariate normal distribution with a mean of 0 and a variance-covariance matrix that allows correlations to exist across choices[30]. The outcome remains as the participants' choice of alternatives for each choice task and the fixed effects predictors are the 16 knowledge translation attributes.

To compare MNL and MNP models, we estimated the normalized covariance matrix of the error difference vector (obtained by differencing each term with respect to the last one in the error vector) in a MNP model. If that covariance matrix is close to the counter-part in a MNL model, which is $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, then a logit model is more appropriate. If values are different from this matrix, then a logit model is not appropriate and a probit model will be used[33].

**2.5 Random-effects MNL and MNP models**

Once we identified the most appropriate fixed effects model (MNL or MNP), we evaluated each with a random-effects model. A hierarchical random effects model was created for random effects MNL and MNP models, unadjusted for potential confounders. The prior distributions for the random effects were set as normal distributions, where the prior mean of the random effects was the mean vector of regression coefficients and the variance followed an inverse Wishart distribution[34].

To create each of the four models, there were some variations in the code specifications for PROC BCHOICE in SAS. A simulation size of 25,000 (coded as nmc in SAS) and a thinning value of 5 were set for random effects only MNP and MNL models; a simulation size of 10,000 and thinning value of 2 were set for the fixed effects MNL model; and a simulation size of 10,000 with thinning of 10 were set for the fixed effects MNP model. This setting yielded an estimated sample size (ESS) of more than 4000 for all parameters in the fixed effects MNL model, more than 500 in the fixed effects MNP model (except for one parameter: the impact on clients will help 33% of clients), and around 500 in both the random effects MNL and MNP models.

**2.6 Ranking of attributes**

The ranking of attributes was determined by transforming the coefficients from each model into measures of relative importance. For each of the four-level attributes, three utility measures are estimated in the regression models (one level is the reference level). To determine the relative importance score for each attribute, the range (difference between the maximum and minimum utility measure) is first determined for each attribute. This range is then divided by the sum of the

ranges of all attributes and multiplied by 100%. Mathematically, for each attribute *k*, the relative importance *(RI)* is determined using the level of attribute *k* with maximum value *($B_{k,max}$)*, level with minimum value *($B_{k,min}$)*, and the total of the absolute range of each attribute:

$$RI_k = \frac{|\beta_{k,max} - \beta_{k,min}|}{\sum_{j=1}^{11} |\beta_{j,max} - \beta_{j,min}|} \; x \; 100\% \quad [21, 35].$$

Each model is presented with mean part-worth utility estimates, standard deviation, 95% HPD Intervals and relative importance (%). Comparisons across models were made using the ranking of attributes. Attribute levels were ranked using the magnitude of each attribute level's mean part-worth utility measure in comparison to the reference level. This ranking was also compared across models.

## 3. RESULTS

Data from 1371 individuals were received from Cunningham and colleagues' DCE, where participants worked in one of 333 different addiction agencies across Canada. A majority of participants were service providers (72%) and others were administrators (26%); more than half were addiction professionals (58%); almost half had more than five years of experience in their position (46%). Many had at least some university level education (66%), an academic affiliation (85%) and worked in a small agency with 1-10 staff (57%). A total of 23,994 choice tasks were completed by participants, with an average of 17.5 completed choice tasks per participant. Incomplete choice tasks were removed before regression analyses. Further participant demographic information is presented in Table 3.

The final model identified for this analysis is the random effects only MNL model. The normalized covariance matrix of the error difference vector of the fixed effects MNP model was $\begin{bmatrix} 1 & 0.49 \\ 0.5 & 1.2 \end{bmatrix}$, allowing an MNL model to be an appropriate model. The DIC was much smaller for the random effects only MNL model (DIC=32797.93) in comparison to the fixed effects MNL model (DIC=35363.18), respectively, making the random effects MNL model better. The estimate of each model's parameter is presented in Table 4.

Each model identified all parameters associated with the outcome except for one. There was no significant difference in the mean part-worth utility between the levels "information is easy to apply" versus the reference level of "information is very easy to apply" in the "implementation complexity" attribute. The mean part-worth utility was -0.03 with an interval containing zero (-0.13, 0.09). This lack of statistical significance was consistent across all models, thus not sensitive to clustering.

The ranking of attributes by relative importance has little sensitivity to clustering effects or heterogeneity within the data. The percentage of relative importance for each attribute was similar across all models. MNL and MNP models identified closer values; however, some variation existed in the ranking of attributes between the fixed effects and random effects only models. For example, the "compatibility with work" attribute ranked higher than "presenter's background" in the fixed effects models and had the opposite order in the random effects only models. More specifically, there was a higher relative importance for "compatibility with work" in the fixed effects MNL (10.21%) and MNP (10.19%) models, than in the random effects only MNL (10.23%) and MNP (10.22%) models. Similarly, the "presenter's background" ranked lower in the fixed

effects MNL (10.09%) and MNP (10.19%) models, while the reverse was found for the random effects only MNL (10.12%) and MNP (10.16%) models. These differences are marginal but they do influence the final ranking of attributes. The relative importance of each attribute in each model is presented in Figure 1.

The top five most preferred attributes across all four models, in the order identified by the random effects MNL model, were “impact on clients” (14.74%), “implementation complexity" (12.83%), “presenter’s background” (10.23%), “compatibility with work” (10.12%), and “collaborative selection process” (8.36%). More specifically, from comparing utility estimates of attribute levels from Table 4, service providers and administrators at addiction agencies for women were most interested in receiving information that would help 100% of their clients, is very easy to apply, and is 100% compatible with their work. They would rather have information presented by researchers than by administrators, former clients, and clinicians. They also preferred to have the responsibility of selecting what content is presented instead of government funders, other administrators, or a team of co-workers including themselves to select the content.

## 4. DISCUSSION

The ranking of attributes’ relative importance measures in DCEs is minimally influenced by the potential clustering of participants’ responses to several choice tasks in a survey. Fixed effects MNL and MNP models and random effects only MNL and MNP models were used to assess the robustness of participant preferences on 16 knowledge translation variables. While the ranking of some attributes differed between fixed effects and random effects models, the differences in relative importance values were marginal. They are likely a consequence of ranking two attributes that essentially have the same value than the use of a (in-)appropriate method. Given

that the two attributes whose ranking differs are only marginally different in relative importance, all models lead to a similar ranking.

The 999 different surveys that were randomly assigned to participants may have also reduced the influence of clustering effects on the ranking. This empirical comparison of methods, which used data from a DCE on 1371 service providers and administrators, identified the five most preferred attributes for knowledge translation were: impact on clients, implementation complexity, presenter's background, compatibility with work, and collaborative selection process.

Cunningham and colleagues used latent class analysis on the same dataset to identify three classes of respondents, outcome sensitive, process sensitive, and demand sensitive segments[27]. They applied a hierarchical Bayes estimation method from Sawtooth Software that determined zero-centered utility coefficients for each attribute and attribute level based on participants within each class [27]. The relative importance ranking of attributes differed by each segment and also differed in comparison to this study's aggregate results. The attribute of highest importance, however, remained consistent across all models and methods. Impact on clients ranked as the highest across the three latent class groups along with the four models we assessed in this study.

To our knowledge, there is limited investigation of the impact of clustering on the robustness of attribute ranking. Cheng and colleagues made similar conclusions in their exploration of ranking robustness across several fixed and random effects models[36]. For a DCE with only two alternatives, choosing between two types of colorectal cancer screening tests, the ranking of relative importance measures were similar across models. In the presence of an "opt out" option,

the rankings had more variability[36]. In their comparisons of the intraclass correlation coefficient between the two datasets, the level of within-participant correlation appears to have an impact in the ranking of the relative importance of attributes. Other papers focus on issues within modelling and introduce approaches to incorporate heterogeneity within DCE data. Flynn and colleagues investigated mean and variance heterogeneity in preferences for quality of life by comparing latent class models with a traditional MNL model[22]. Their focus, however, was to introduce an approach for modelling variance heterogeneity[22]. Keane discussed the issues of using a MNL logit model to make strong assumptions of consumer behaviour for discrete choice data [37]. The MNL logit model assumes homogeneity of the intercept and slope parameters in the population, ignoring heterogeneity in preferences, and ultimately not capturing the differing error variance across choice tasks (since the error variance is assumed as constant for MNL models), and not capturing the serially correlated errors resulting from preference heterogeneity. Keane advised readers to incorporate a heterogeneity structure when modelling DCE data, since modelling heterogeneity will help determine accurate estimates of parameters[37]. Instead of only finding an appropriate statistical model for a research question, investigators should also first consider an appropriate theoretical model that best describes participant behaviours and then derive the statistical model. This will merge the investigator's understanding of the participant group with an appropriate statistical analysis[37]. Stone and Rasp performed a simulation study along with an empirical comparison where they explore the impact of sample size on parameter estimates[38]. Comparing estimates derived from an ordinary least squares linear probability model and a logit model, they found that logit test statistics were biased for smaller sample sizes. They suggested a rule of thumb of 50 cases per parameter. In our study, there were 16 four-level variables, which made for 48 parameters (16*3; since no parameter estimates were needed for reference levels)

which means this study requires 2400 cases. We had a total of 23994 complete cases, making our sample size more than adequate for analysis. They also mentioned that increasing the sample size decreases the difference between nominal and empirical error rates. For studies with a sample size of less than 100, they found that it did not make a difference whether an ordinary least squares model or a probit model was used. Another study, by Signorino, theoretically explored potential sources of uncertainty in structural assumptions when producing different statistical models and their potential influence on inferences[39]. Lastly, Horowitz explored a Lagrangian multiplier test to compare and determine if a logit or probit model is more appropriate given a specific DCE dataset[40]. No studies, from our search, compared the robustness of ranking attributes.

This study has several strengths. Firstly, we identified the high variability in methods to analyze DCEs from literature and systematic reviews. By performing an empirical comparison of methods, where two adjusted for clustering and two did not, we were able to assess whether this variability should be of concern given the possible heterogeneity from repeated measures. Secondly, we explored the use of Bayesian MNL logit and probit models to estimate part-worth utility measures in DCEs. These models were created using PROC BCHOICE, a procedure that was recently developed in SAS (Version 9.4). Thirdly, we identified qualities of how to best package evidence-based findings that are appropriate and of more interest to individuals working at addiction agencies for women. More specifically, we specified the relative importance of 16 different attributes, ranking them in order from most preferred to least preferred. We also mentioned the most favourable level within each of the top five attributes.

Despite our key findings, more could be done to improve this study. The robustness of the rankings is only tested using a single empirical dataset and hence we cannot be sure that it also holds in different DCE settings (e.g. smaller sample size, number of choice tasks, amount of actual heterogeneity, etc.). Participant demographics were not explored as potential confounders. Since the models already had 48 parameters, adding demographics to the PROC BCHOICE method in SAS is a complex process, especially for fixed effects models. Interactions between attributes or attribute levels were also not explored.

Further analyses could be performed to overcome these limitations. First, simulation studies and other analyses on empirical datasets should be conducted in order to further address the aim of the study in various DCE design settings. This is to ensure the findings are generalizable in different situations (e.g. smaller sample size, number of choice tasks, amount of actual heterogeneity, etc.). Second, the impact of potential confounders, such as participant demographics, should be explored in models to see if their presence reduces the sensitivity in the ranking of attributes across models. This was not discussed in the reviews mentioned in Table 1. Third, a systematic review of methods in applications of DCEs would provide more insight into the use and reporting of software and models performed, criteria used to determine final models, goodness of fit measures, etc.

## 5. CONCLUSIONS

Our findings suggest that, in the study we considered, there is minimal effect of heterogeneity on the ranking of attributes in discrete choice experiments. Whether these results also apply in general in different settings of DCEs should be further investigated. Both random effects and fixed

effects approaches were used for MNL and MNP models, where all reported similar mean part-worth utility measures and relative importance (%) measures. We advise investigators who are identifying the top three or top five most preferred attributes to proceed with caution when the consecutively ranked attribute only differs by a small amount. This may have a slightly higher rank depending on the model used for analysis. As for individuals interested in disseminating information into addiction agencies for women, it would be most beneficial to consider information that will help the majority of clients, will be very easy to apply, will be largely compatible with the service providers' work, and provide them with the responsibility of selecting what content is presented. Further research should include other empirical studies and simulation studies, with varying participant demographics and design characteristics, to ensure these conclusions are generalizable to other DCE settings.

## 6. ADDITIONAL DETAILS

### Author Contributions (described using authors' initials)

1) Substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data:

   TV, LT, CC, GF, AS, HR, JH, AN, MD, WS, YC, SM, KM, EL, and LS

2) Drafting the article or revising it critically for important intellectual content:

   TV, LT, CC, GF, AS, HR, JH, AN, MD, WS, YC, SM, KM, EL, and LS

3) Final approval of the version to be published:

   TV, LT, CC, GF, AS, HR, JH, AN, MD, WS, YC, SM, KM, EL, and LS

4) Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and

resolved:

TV, LT, CC, GF, AS, HR, JH, AN, MD, WS, YC, SM, KM, EL, and LS

**Data sharing:**

Statistical code is available upon request to the authors

**Conflict of Interest:**

TV, LT, GF, AS, HR, JH, AN, MD, WS, YC, SM, KM, EL, and LS declare: no conflicts of interest. CC's participation was supported by the Jack Laidlaw Chair in Patient-Centered Health Care. The project that had collected the original dataset was supported by a grant from the Canadian Institutes of Health Research (CIHR).

**Funding sources:**

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**TABLES AND FIGURES:**

Table 1: DCE designs and statistical methods reported by systematic reviews of DCEs

| **Description of Systematic Reviews** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **First author [Reference]** | Ryan [12] | de Bekker-Grob [15] | Mandeville [14] | Regmi [16] | Michaels-Igbokwe [18] | de Bekker-Grob [15] | Clark [23] | Vass [17] |
| **Year Published** | 2003 | 2012 | 2014 | 2018 | 2017 | 2012 | 2014 | 2016 |
| **Years covered** | 1990-2000 | 1990-2000 | 1998-2013 | 2000-2016 | 2000-2016 | 2001-2008 | 2009-2012 | 2012-2015 |
| **Total # studies assessed** | 34 | 34 | 27 | 12 | 27 articles representing 21 different studies | 114 | 179 | 17 |
| **Design and Analyses of Studies** | | | | | | | | |
| **Design types** | Full factorial (4)<br>Fractional factorial (25)<br>Not reported (5) | Full factorial (4)<br>Fractional factorial (25)<br>Unclear (5) | Main effects only (4)<br>Main effects with interaction (1)<br>not reported/ unclear (22) | Fractional factorial (9)<br>Factorial(1)<br>Balanced (1)<br>Balance incomplete block (1) | Full factorial (1)<br>Fractional factorial (15)<br>Not reported (5) | Fractional factorial (114) | *Percentages were obtained from graphs:*<br>Full factorial (6%)<br>Fractional factorial (88%)<br>Main effects (54%)<br>Main effects and interactions (13%)<br>Not applicable (<5%)<br>Not reported (7%) | Fractional Factorial (16)<br>Unclear/not reported(1) |
| **Statistical Model** | RE probit (17)<br>Logit/MNL (9)<br>Probit/ordered probit (3)<br>Other (3)<br>Not reported (2) | RE probit (18)<br>Probit (6)<br>MNL (6)<br>RE logit (1)<br>Logit (1)<br>Mixed logit (1)<br>Other (1)<br>Unclear(2) | Mixed logit (11)<br>RE probit (7)<br>Generalized MNL (4)<br>Conditional logit (3)<br>Logit (2)<br>Probit (1)<br>MNL (1)<br>Errors component mixed logit(1) | MNL logit(3)<br>Conditional logit(2)<br>Random parameter logit(2)<br>Mixed logit(2)<br>Nested logit(1)<br>RE probit (1)<br>Other (2) | Conditional/ MNL logit (6)<br>Mixed logit (6)<br>RE/mixed logit (4)<br>Hierarchical Bayes estimation (3)<br>Generalized mixed logit(1)<br>Latent class (2)<br>Other (2) | RE probit (47)<br>MNL (25)<br>Logit (13)<br>Probit (8)<br>RE logit (6)<br>Nested logit (5)<br>Mixed logit (6)<br>Latent class (1)<br>Other (4)<br>Unclear(4) | MNL (44%)<br>Mixed logit or random parameter logit (RPL) (21%)<br>RE probit (10%)<br>Logit (10%)<br>RE Logit (<10%)<br>Latent class (3%)<br>Nested logit (2%)<br>Probit (2%)<br>Other (17)<br>Not reported (<5%) | Conditional logit/MNL logit (10)<br>RE probit (3)<br>Heteroskedastic MNL logit (2)<br>Latent-class analysis (2)<br>Mixed logit (2)<br>RE conditional logit (1)<br>Other (2) |

**RE=Random effects; MNL=multinomial logit*

Table 2: Knowledge translation variables under investigation

| Area | Attributes | Levels |
|---|---|---|
| Expected Outcome | Impact on clients | Will help 0% of my clients<br>Will help 33% of my clients<br>Will help 67% of my clients<br>Will help 100% of my clients |
| | Quality of evidence | None<br>Staff from other agencies<br>Research<br>Research and staff from other agencies |
| Feasibility | Compatibility | 0% compatible with my work<br>33% compatible with my work<br>67% compatible with my work<br>100% compatible with my work |
| | Implementation complexity | Information is very easy to apply<br>Information is easy to apply<br>Information is difficult to apply<br>Information is very difficult to apply |
| | Number of days for training (Time cost) | Requires 1 day of my time<br>Requires 2 days of my time<br>Requires 3 days of my time<br>Requires 4 days of my time |
| Support | Co-worker support | Supported by 0% of my co-workers<br>Supported by 33% of my co-workers<br>Supported by 67% of my co-workers<br>Supported by 100% of my co-workers |
| | Administrative support | Discouraged by my boss<br>My boss is neutral<br>Supported by my boss<br>Strongly supported by my boss |
| | Source of endorsement | Not endorsed<br>Endorsed by the government<br>Endorsed by a colleague<br>Endorsed by an expert |
| | Collaborative selection process | Government funders select the content<br>Administrators select the content<br>I select the content<br>My co-workers and I select the content |

| | | |
|---|---|---|
| Implementation Process | Presenter's background | Presenter is an administrator<br>Presenter is a former client<br>Presenter is a researcher<br>Presenter is a clinician |
| | Accessing supplementary information | More information is not available<br>More information is on a website<br>More information is sent if I request it<br>More information is automatically sent |
| | Internet options | Involves no electronic media<br>Involves a blog, list serve, twitter, or face book<br>Involves a web seminar<br>Involves a self-paced internet program |
| | Focus on knowledge versus skills | 100% focus on knowledge<br>67% focus on knowledge, 33% on skills<br>33% focus on knowledge, 67% on skills<br>100% focus on skills |
| | Individuals versus group format | Includes no 1:1 contact or workshops<br>Includes 1:1 contact<br>Includes a small group workshop<br>Includes a large group workshop |
| | Active versus passive-learning | No review questions or practice exercises<br>Includes review questions<br>Includes practice exercises<br>Review questions & practice exercises |
| | Number of implementation follow-ups | Includes 0 implementation follow-ups<br>Includes 1 implementation follow-up<br>Includes 2 implementation follow-ups<br>Includes 3 implementation follow-ups |

Table 3: Participant Demographics

| Participant Characteristic (n=1371) | n (%) |
|---|---|
| **Job Type** | |
| Administrator | 353(25.7) |
| Service provider | 980(71.5) |
| **Addiction Professional** | |
| No | 542(39.5) |
| Yes | 794(57.9) |
| **Years in Position** | |
| <2 years | 319(23.3) |
| 2-5 years | 395(28.8) |
| >5years | 626(45.7) |
| **Level of Education** | |
| High school | 55(4.0) |
| College | 361(26.3) |
| Any university | 909(66.3) |
| **Agency Location** | |
| Not urban | 225(16.4) |
| Urban | 1112(81.1) |
| **Agency Size** | |
| Small (1-10 staff) | 783(57.1) |
| Medium (11-20 staff) | 276(20.1) |
| Large (20+ staff) | 276(20.1) |
| **Academic Affiliation** | |
| No | 169(12.3) |
| Yes | 1166(85.0) |

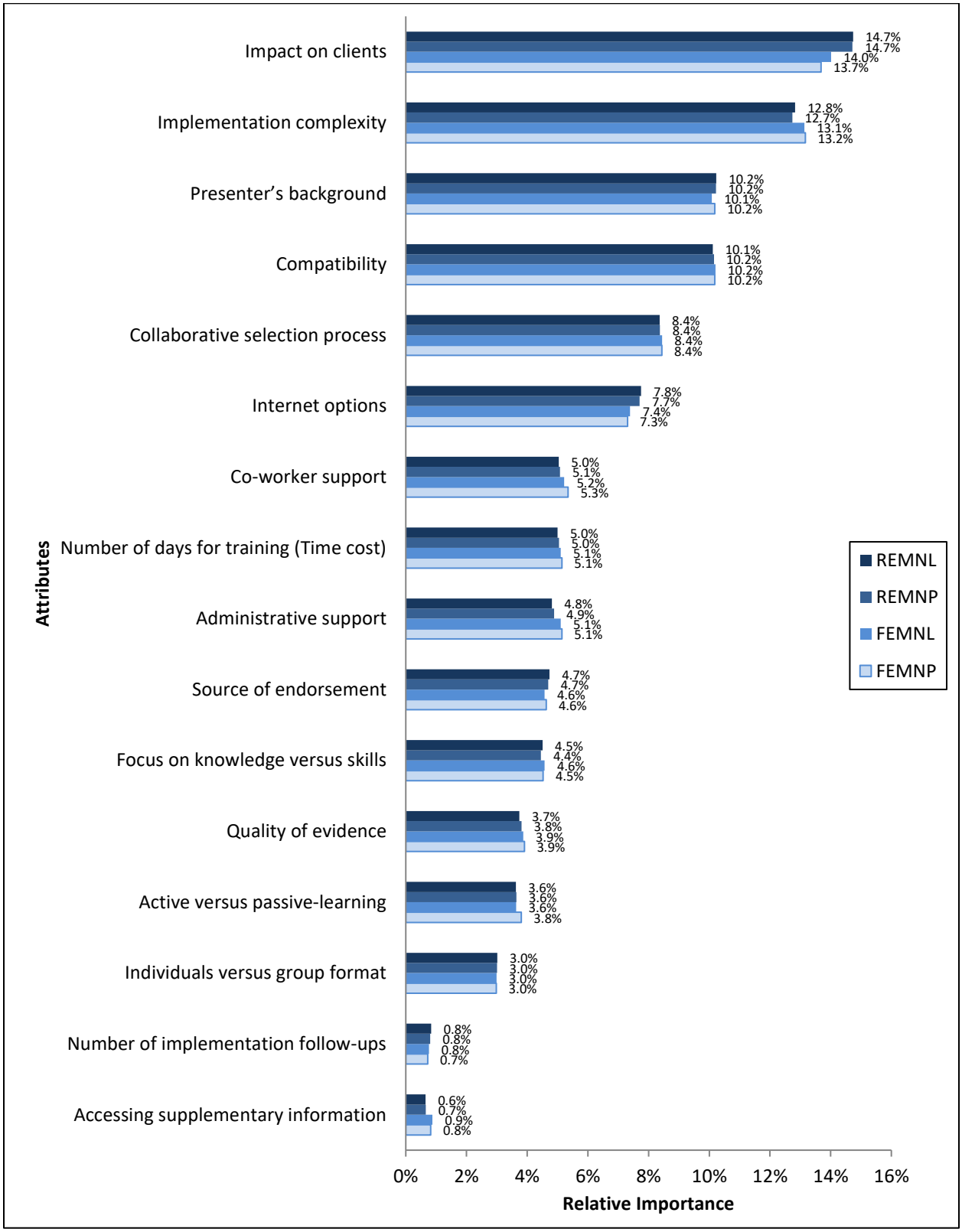
Table 4: Mean part-worth value of participant preferences for fixed effects and random effects MNL and MNP models

| Attribute | Attribute Level | Fixed effects MNL | Fixed effects MNP | Random Effects MNL | Random Effects MNP |
|---|---|---|---|---|---|
| | | Mean (95% HPDI) | Mean (95% HPDI) | Mean (95% HPDI) | Mean (95% HPDI) |
| **Quality of evidence** | None | REF | | . | . |
| | Staff from other agencies | 2.2 (2.1, 2.4) | 1.2 (1.1, 1.3) | 3.1 (2.9, 3.3) | 1.7 (1.5, 1.8) |
| | Research | 2.3 (2.1, 2.4) | 1.2 (1.1, 1.3) | 3.1 (2.9, 3.3) | 1.7 (1.5, 1.8) |
| | Research and staff from other agencies | 2.9 (2.7, 3.0) | 1.6 (1.5, 1.7) | 4.1 (3.8, 4.3) | 2.2 (2.0, 2.4) |
| **Co-worker support** | Supported by 0% of my co-workers | REF | . | . | . |
| | Supported by 33% of my co-workers | 1.3 (1.2, 1.5) | 0.7 (0.7, 0.8) | 1.8 (1.6, 2.0) | 1.0 (0.9, 1.1) |
| | Supported by 67% of my co-workers | 2.0 (1.8, 2.1) | 1.1 (1.0, 1.2) | 2.7 (2.5, 2.9) | 1.5 (1.4, 1.6) |
| | Supported by 100% of my co-workers | 2.2 (2.1, 2.4) | 1.3 (1.2, 1.3) | 3.2 (3.0, 3.4) | 1.7 (1.6, 1.9) |
| **Administrative support** | Discouraged by my boss | REF | . | . | . |
| | My boss is neutral | 1.4 (1.2, 1.5) | 0.8 (0.7, 0.8) | 1.9 (1.7, 2.1) | 1.0 (0.9, 1.2) |
| | Supported by my boss | 2.1 (1.9, 2.2) | 1.2 (1.1, 1.2) | 2.9 (2.7, 3.1) | 1.6 (1.4, 1.7) |
| | Strongly supported by my boss | 2.3 (2.1, 2.4) | 1.3 (1.2, 1.3) | 3.2 (3.0, 3.4) | 1.7 (1.6, 1.9) |
| **Individuals versus group format** | Includes no 1:1 contact or workshops | REF | . | . | . |
| | Includes 1:1 contact | 0.9 (0.8, 1.0) | 0.5 (0.4, 0.6) | 1.2 (1.0, 1.4) | 0.7 (0.6, 0.8) |
| | Includes a small group workshop | 1.4 (1.3, 1.5) | 0.8 (0.7, 0.9) | 2.0 (1.8, 2.2) | 1.1 (1.0, 1.2) |
| | Includes a large group workshop | 0.9 (0.8, 1.1) | 0.5 (0.5, 0.6) | 1.3 (1.1, 1.5) | 0.7 (0.6, 0.8) |
| **Internet options** | Involves no electronic media | REF | . | . | . |
| | Involves a blog, list serve, twitter, or face book | 0.5 (0.4, 0.6) | 0.3 (0.2, 0.4) | 0.7 (0.6, 0.9) | 0.4 (0.3, 0.5) |
| | Involves a web seminar | 0.2 (0.1, 0.3) | 0.1 (0.0, 0.2) | 0.2(0.0, 0.4) | 0.1 (0.0, 0.2) |
| | Involves a self-paced internet program | -0.7 (-0.9, -0.6) | -0.4 (-0.5, -0.4) | -1.3 (-1.5, -1.1) | -0.7 (-0.8, -0.6) |
| **Active versus passive-learning** | No review questions or practice exercises | REF | . | . | . |
| | Includes review questions | 0.6 (0.5, 0.8) | 0.4 (0.3, 0.4) | 0.8 (0.6, 1.0) | 0.5 (0.4, 0.6) |
| | Includes practice exercises | 1.1 (0.9, 1.2) | 0.6 (0.5, 0.7) | 1.4 (1.3, 1.6) | 0.8 (0.7, 0.9) |
| | Review questions & practice exercises | 1.3 (1.1, 1.4) | 0.7 (0.6, 0.8) | 1.8 (1.6, 1.9) | 1.0 (0.9, 1.1) |
| **Presenter's background** | Presenter is an administrator | REF | . | . | . |
| | Presenter is a former client | 0.4 (0.2, 0.5) | 0.2 (0.1, 0.3) | 0.5 (0.3, 0.7) | 0.3 (0.2, 0.3) |
| | Presenter is a researcher | 1.0 (0.9, 1.1) | 0.6 (0.5, 0.7) | 1.5 (1.3, 1.7) | 0.8 (0.7, 0.9) |
| | Presenter is a clinician | -0.7 (-0.8, -0.6) | -0.4 (-0.5, -0.3) | -1.2 (-1.4, -1.0) | -0.6(-0.8, -0.5) |
| **Number of implementation follow-ups** | Includes 0 implementation follow-ups | REF | . | . | . |
| | Includes 1 implementation follow-up | 0.8 (0.6, 0.9) | 0.4 (0.4, 0.5) | 1.0 (0.9, 1.2) | 0.6 (0.5, 0.7) |
| | Includes 2 implementation follow-ups | 0.8 (0.7, 0.9) | 0.5 (0.4, 0.5) | 1.1 (0.9, 1.3) | 0.6 (0.5, 0.7) |
| | Includes 3 implementation follow-ups | 0.7 (0.6, 0.8) | 0.4 (0.3, 0.4) | 0.9 (0.7, 1.1) | 0.5 (0.4, 0.6) |
| **Focus on knowledge versus skills** | 100% focus on knowledge | REF | . | . | . |
| | 67% focus on knowledge, 33% on skills | 1.2 (1.1, 1.3) | 0.7 (0.6, 0.7) | 1.6 (1.5, 1.8) | 0.9 (0.8, 1.0) |
| | 33% focus on knowledge, 67% on skills | 1.4 (1.3, 1.6) | 0.8 (0.7, 0.9) | 2.0 (1.8, 2.2) | 1.1 (1.0, 1.2) |
| | 100% focus on skills | 0.7 (0.5, 0.8) | 0.4 (0.3, 0.4) | 0.8 (0.7, 1.0) | 0.5 (0.4, 0.6) |
| | Government funders select the content | REF | . | . | . |
| | Administrators select the content | 0.3 (0.2, 0.5) | 0.2 (0.1, 0.3) | 0.4 (0.2, 0.6) | 0.2 (0.1, 0.3) |

| | | | | | |
|---|---|---|---|---|---|
| **Collaborative selection process** | I select the content | 1.8 (1.6, 1.9) | 1.0 (0.9, 1.1) | 2.6 (2.4, 2.8) | 1.4 (1.3, 1.5) |
| | My co-workers and I select the content | 1.1 (1.0, 1.3) | 0.6 (0.6, 0.7) | 1.6 (1.4, 1.8) | 0.9 (0.8, 1.0) |
| **Implementation complexity** | Information is very easy to apply | REF | . | . | . |
| | Information is easy to apply | <0.01 (-0.1, 0.1) | <0.01 (-0.1, 0.1) | <0.01 (-0.2, 0.1) | <0.01 (-0.1, 0.1) |
| | Information is difficult to apply | -1.8 (-1.9, -1.7) | -1.0 (-1.1, -0.9) | -2.7 (-2.9, -2.5) | -1.5 (-1.6, -1.3) |
| | Information is very difficult to apply | -2.3 (-2.4, -2.1) | -1.3 (-1.4, -1.2) | -3.4 (-3.6, -3.2) | -1.8 (-2.0, -1.7) |
| **Impact on clients** | Will help 0% of my clients | REF | . | . | . |
| | Will help 33% of my clients | 3.1 (2.9, 3.4) | 1.6 (1.4, 1.7) | 4.2 (3.8, 4.5) | 2.2 (2.0, 2.4) |
| | Will help 67% of my clients | 4.5 (4.2, 4.8) | 2.3 (2.2, 2.5) | 6.3 (6.0, 6.8) | 3.4 (3.1, 3.6) |
| | Will help 100% of my clients | 5.5 (5.2, 5.8) | 2.9 (2.8, 3.0) | 8.0 (7.6, 8.5) | 4.3 (3.9, 4.6) |
| **Number of days for training (Time cost)** | Requires 1 day of my time | REF | . | . | . |
| | Requires 2 days of my time | -0.2 (-0.3, -0.1) | -0.1 (-0.2, -0.1) | -0.3 (-0.4, -0.1) | -0.2 (-0.3, -0.1) |
| | Requires 3 days of my time | -0.6 (-0.7, -0.5) | -0.3 (-0.4, -0.3) | -0.9 (-1.0, -0.7) | -0.5 (-0.6, -0.4) |
| | Requires 4 days of my time | -1.1 (-1.2, -0.9) | -0.6 (-0.7, -0.5) | -1.6 (-1.8, -1.4) | -0.9 (-1.0, -0.8) |
| **Accessing supplementary information** | More information is not available | REF | . | . | . |
| | More information is on a website | 1.5 (1.4, 1.7) | 0.9 (0.8, 0.9) | 2.2 (2.0, 2.4) | 1.2 (1.1, 1.3) |
| | More information is sent if I request it | 1.4 (1.3, 1.6) | 0.8 (0.7, 0.9) | 2.0 (1.8, 2.2) | 1.1 (1.0, 1.2) |
| | More information is automatically sent | 1.6 (1.4, 1.7) | 0.9 (0.8, 1.0) | 2.2 (2.1, 2.4) | 1.2 (1.1, 1.3) |
| **Compatibility** | 0% compatible with my work | REF | . | . | . |
| | 33% compatible with my work | 1.9 (1.7, 2.1) | 1.0 (0.9, 1.1) | 2.5 (2.3, 2.7) | 1.3 (1.2, 1.5) |
| | 67% compatible with my work | 3.0 (2.8, 3.2) | 1.6 (1.5, 1.7) | 4.2 (4.0, 4.4) | 2.3 (2.1, 2.5) |
| | 100% compatible with my work | 3.6 (3.5, 3.8) | 2.0 (1.9, 2.1) | 5.2 (4.9, 5.4) | 2.8 (2.6, 3.0) |
| **Source of endorsement** | Not endorsed | REF | . | . | . |
| | Endorsed by the government | 1.3 (1.1, 1.4) | 0.7 (0.6, 0.8) | 1.8 (1.6, 2.0) | 1.0 (0.9, 1.1) |
| | Endorsed by a colleague | 1.8 (1.7, 1.9) | 1.0 (0.9, 1.1) | 2.6 (2.4, 2.8) | 1.4 (1.3, 1.6) |
| | Endorsed by an expert | 1.0 (0.9, 1.1) | 0.6 (0.5, 0.6) | 1.4 (1.2, 1.6) | 0.8 (0.6, 0.9) |

Comment: MNL=multinomial logit model; MNP= multinomial probit model; HPDI=highest posterior density interval; REF=reference level.

Figure 1: Relative importance of attributes for each model




Comment: REMNL=Random effects multinomial logit model; REMNP=random effects multinomial probit model; FEMNL=fixed effects multinomial logit model; FEMNP= fixed effects multinomial probit model

**REFERENCES**

1. Sepucha K, Ozanne EM: **How to define and measure concordance between patients' preferences and medical treatments: A systematic review of approaches and recommendations for standardization**. *Patient Educ Couns* 2010, **78**(1):12-23.
2. Pfarr C, Schmid A, Schneider U: **Using discrete choice experiments to understand preferences in health care**. *Dev Health Econ Public Policy* 2014, **12**:27-48.
3. Ryan M: **Discrete choice experiments in health care**. *BMJ (Clinical research ed)* 2004, **328**(7436):360-361.
4. Wong SF, Norman R, Dunning TL, Ashley DM, Khasraw M, Hayes TM, Collins I, PK. L: **A Discrete Choice Experiment to Examine the Preferences of Patients With Cancer and Their Willingness to Pay for Different Types of Health Care Appointments**. *J Natl Compr Canc Netw* 2016, **14**(3):311-319. Epub 2016 Mar 2018.
5. Rowen D, Brazier J, Mukuria C, Keetharuth A, Risa Hole A, Tsuchiya A, Whyte S, Shackley P: **Eliciting Societal Preferences for Weighting QALYs for Burden of Illness and End of Life**. *Medical decision making : an international journal of the Society for Medical Decision Making* 2016, **36**(2):210-222.
6. Grindrod KA, Marra CA, Colley L, Tsuyuki RT, Lynd LD: **Pharmacists' preferences for providing patient-centered services: a discrete choice experiment to guide health policy**. *The Annals of pharmacotherapy* 2010, **44**(10):1554-1564.
7. Koopmanschap MA, Stolk EA, Koolman X: **Dear policy maker: have you made up your mind? A discrete choice experiment among policy makers and other health professionals**. *International journal of technology assessment in health care* 2010, **26**(2):198-204.

8. Berhane A, Enquselassie F: **Patients' preferences for attributes related to health care services at hospitals in Amhara Region, northern Ethiopia: a discrete choice experiment**. *Patient Prefer Adherence* 2015, **9**:1293-1301.

9. Gray E, Eden M, Vass C, McAllister M, Louviere J, Payne K: **Valuing Preferences for the Process and Outcomes of Clinical Genetics Services: A Pilot Study**. *Patient* 2016, **9**(2):135-147.

10. Arellano J, Gonzalez JM, Qian Y, Habib M, Mohamed AF, Gatta F, Hauber AB, Posner J, Califaretti N, Chow E: **Physician preferences for bone metastasis drug therapy in Canada**. *Current oncology (Toronto, Ont)* 2015, **22**(5):e342-348.

11. Berchi C, Degieux P, Halhol H, Danel B, Bennani M, Philippe C: **Impact of falling reimbursement rates on physician preferences regarding drug therapy for osteoarthritis using a discrete choice experiment**. *The International journal of pharmacy practice* 2016, **24**(2):114-122.

12. Ryan M, Gerard K: **Using discrete choice experiments to value health care programmes: current practice and future research reflections**. *Applied health economics and health policy* 2003, **2**(1):55-64.

13. Marshall D, Bridges JF, Hauber B, Cameron R, Donnalley L, Fyie K, Johnson FR: **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**(4):249-256.

14. Mandeville KL, Lagarde M, Hanson K: **The use of discrete choice experiments to inform health workforce policy: a systematic review**. *BMC Health Services Research* 2014, **14**(1):367.

15. de Bekker-Grob EW, Ryan M, Gerard K: **Discrete choice experiments in health economics: a review of the literature**. *Health economics* 2012, **21**(2):145-172.

16. Regmi K, Kaphle D, Timilsina S, Tuha NAA: **Application of Discrete-Choice Experiment Methods in Tobacco Control: A Systematic Review**. *PharmacoEconomics Open* 2018, **2**(1):5-17.

17. Vass C, Gray E, Payne K: **Discrete choice experiments of pharmacy services: a systematic review**. *International journal of clinical pharmacy* 2016, **38**(3):620-630.

18. Michaels-Igbokwe C, MacDonald S, Currie GR: **Individual Preferences for Child and Adolescent Vaccine Attributes: A Systematic Review of the Stated Preference Literature**. *Patient* 2017, **10**(6):687-700.

19. Lancaster KJ: **A new approach to consumer theory**. *The journal of political economy* 1966:132-157.

20. Muhlbacher A, Johnson FR: **Choice Experiments to Quantify Preferences for Health and Healthcare: State of the Practice**. *Appl Health Econ Health Policy* 2016.

21. Orme B: **Interpreting the Results of Conjoint Analysis**. In: *Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research, Second Edition.* edn. Madison, Wis: Research Publishers, LLC; 2010: 77-89.

22. Flynn TN, Louviere JJ, Peters TJ, Coast J: **Using discrete choice experiments to understand preferences for quality of life. Variance-scale heterogeneity matters**. *Social science & medicine (1982)* 2010, **70**(12):1957-1965.

23. 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.

24. Li B, Lingsma HF, Steyerberg EW, Lesaffre E: **Logistic random effects regression models: a comparison of statistical packages for binary and ordinal outcomes**. *BMC medical research methodology* 2011, **11**(1):1-11.

25. Hensher DA, Greene WH: **The mixed logit model: the state of practice**. *Transportation* 2003, **30**(2):133-176.

26. Hensher DA, Jones S: **Mixed logit and error component model of corporate insolvency and bankruptcy risk**. *Advances in Credit Risk Modelling and Corporate Bankruptcy Prediction* 2008:44.

27. Cunningham CE, Henderson J, Niccols A, Dobbins M, Sword W, Chen Y, Mielko S, Milligan K, Lipman E, Thabane L *et al*: **Preferences for evidence-based practice dissemination in addiction agencies serving women: a discrete-choice conjoint experiment**. *Addiction* 2012, **107**(8):1512-1524.

28. Stokes M, Chen F, Gunes F: **An introduction to Bayesian analysis with SAS/STAT® software**. In: *Proceedings of the SAS Global Forum 2014 Conference, SAS Institute Inc, Cary, USA (available at https://supportsas com/resources/papers/proceedings14/SAS400-2014 pdf): 2014*: Citeseer; 2014.

29. McDowell A, Shi A: **Introducing the BCHOICE Procedure for Bayesian Discrete Choice Models**. 2014.

30. Vojáček O, Pecáková I: **Comparison of discrete choice models for economic environmental research**. *Prague Economic Papers* 2010, **19**(1):35-53.

31. Spiegelhalter DJ, Best NG, Carlin BP, Linde A: **The deviance information criterion: 12 years on**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2014, **76**(3):485-493.

32. Hanley N, Mourato S, Wright RE: **Choice Modelling Approaches: A Superior Alternative for Environmental Valuatioin?** *Journal of economic surveys* 2001, **15**(3):435-462.

33. SAS Institute Inc: **The BCHOICE Procedure**. In: *SAS/STAT® 141 User's Guide.* Cary, NC: SAS Institute Inc.; 2015: 1025-1105.

34. McDowell A, Shi A: **Introducing the BCHOICE Procedure for Bayesian Discrete Choice Models**. In.: SAS Institute Inc.; 2014: 1-20.

35. Sullivan LM, Massaro JM, D'Agostino RB: **Presentation of multivariate data for clinical use: the Framingham Study risk score functions**. *Statistics in medicine* 2004, **23**.

36. Cheng J, Pullenayegum E, Marshall DA, Marshall JK, Thabane L: **An empirical comparison of methods for analyzing correlated data from a discrete choice survey to elicit patient preference for colorectal cancer screening**. *BMC medical research methodology* 2012, **12**(1):1.

37. Keane M: **Current issues in discrete choice modeling**. *Marketing Letters* 1997, **8**(3):307-322.

38. Stone M, Rasp J: **Tradeoffs in the choice between logit and OLS for accounting choice studies**. *Accounting review* 1991:170-187.

39. Signorino CS: **Structure and uncertainty in discrete choice models**. *Political Analysis* 2003, **11**(4):316-344.

40. Horowitz J: **Testing the multinomial logit model against the multinomial probit model without estimating the probit parameters**. *Transportation Science* 1981, **15**(2):153-163.

# CHAPTER 5

## Investigating the impact of clustering effects through hierarchical models for analyzing discrete choice experiments: an empirical comparison

Thuva Vanniyasingam[1,2]*, Charles Cunningham[3], Heather Rimas[3], Bailey Stewart[3], Gary r

[1]Department of Health Research Methods, Impact, and Evidence, McMaster University, Hamilton, ON, Canada
[2]Biostatistics Unit, Father Sean O'Sullivan Research Centre, St. Joseph's Healthcare, Hamilton, ON, Canada.
[3]Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, ON, Canada
[4]Departments of Paediatrics and Anaesthesia, McMaster University, Hamilton, ON, Canada
[5]Centre for Evaluation of Medicine, St. Joseph's Healthcare, Hamilton, ON, Canada
[6]Population Health Research Institute, Hamilton Health Sciences, Hamilton, ON, Canada

***Corresponding Author:**

Thuva Vanniyasingam

Department of Clinical Epidemiology and Biostatistics, HSC 2C

McMaster University, Hamilton, ON, L8S 4L8

thuva.vanni@gmail.com

**Abstract**

**Background**

Discrete choice experiments (DCEs) are a tool in health research for eliciting individuals' preferences on various attributes describing a health care product or service. The ranking of these preferences relies on participants' responses to several questions in a survey, leading to an issue of multiple response. It is unclear how these correlated responses and hierarchical data structures impact the final ranking of these attributes in DCEs.

**Objective**

The aim of this study was to determine how robust the rankings of attributes' re measures were across various hierarchical model settings, using an empirical survey dataset.

**Methods**

A survey was conducted with elementary school children who were presented with schools containing various alternatives of antibullying programs. The survey sought to determine their choice of school where students were more likely to speak up, report bullying, or seek help from an adult (in response to bullying). A partial profile DCE design was used to investigate the impact of 11 attributes (e.g. reporting by peers), each with three levels (e.g. almost never, sometimes, almost always), on participants' choice of anti-bullying programs in schools. Nine models were created: one was a fixed-effects conditional logit (CL) model and eight were random-effects CL models exploring the impact of individuals, classroom, grade, and school as potential clustering variables. The clustering variables were explored using latent class analyses for the individual level, group-level latent class analyses for the classroom and school levels, and using a covariate for the grade level. Each model's mean part-worth utility estimates were used to derive the relative

importance measure for each attribute. The mean part-worth utility estimates within each attribute were then ordered to determine the ranking of attribute levels.

**Results**

A total of 2033 participants were included. Participants were from four grades (5,6,7, and 8), 116 classrooms, and 30 schools. The rankings of relative importance measures were similar across all models. The school anti-bullying program attributes ranked most important were: promptness in response to reports of bullying, the inclusion of students who were left out, and the interest level of anti-bullying presentations. The rankings of levels within attributes were also robust across all models.

**Conclusions**

The rankings of attributes and attribute levels were relatively robust across the various hierarchical models adjusting for clustering. Further research should explore the impact of various levels of correlated data (e.g. similarities in responses among individuals, classrooms, grades, and schools) and their clustering effect on the ranking of attributes using a simulation study.

**Key words:** discrete choice experiment, latent class, participant preference, empirical comparison

## 1. Introduction

Cluster-correlated data arise when there is a grouped structure to a dataset. These groups (clusters) are multi-leveled, or hierarchical, and result in more similar observations within a cluster than across other clusters[1]. This is either due to a naturally occurring hierarchy in the study population (e.g. multiple sites), a consequence of the study design (e.g. repeated measures), or both. An example of a cluster-correlated data structure would be a clinical trial where blood samples (level 1) are measured repeatedly in the same patients (level 2), who are from one of several study sites (level 3).

Discrete choice experiments (DCEs) also have a naturally occurring multilevel data structure. They are a quantitative technique for eliciting preferences that are otherwise unmeasurable. Participants are surveyed on their choice of products, services, or programs, which are characterized by different attributes. Each survey question (choice tasks) presents hypothetical alternatives of various combinations of attributes and attribute levels. As each participant responds to several choice tasks within the survey, a naturally occurring two-level data structure arises. Each response to a choice task (level 1) is nested within participants (level 2), and based on the lived experiences of each participant, their responses to the several choice tasks within a survey may be similar. Participants may also belong to groups (level 3), such as sharing a similar demographic or geographical location.

The overall goal of DCEs is to quantify participants' preferences of attributes describing a product, service, or program. Responses to choice tasks are incorporated into a regression model to produce regression coefficients, referred to as mean part-worth utility estimates (or part-worth

utility estimates, depending on the model). In non-DCE studies, we typically look at the size and statistical significance of the regression coefficients to determine the impact of independent variables on the outcome. In DCEs, however, we aim to compare between these independent variables (attributes) to determine participants' relative preference of an attribute in comparison to the others in the study[2].

The regression coefficients of statistical models analyzing DCE data cannot be used to directly compare preferences of attributes. This is because attributes may differ in origin and units[3]. For example, time and price attributes differ in units by hours and dollars, respectively. Qualitative or nominal attributes also do not have a meaningful scale to make comparisons with other attributes [3]. We can derive relative importance measures using the regression coefficients (or utility estimates) to present individuals' preferences on a common scale[4]. These are calculated by transforming the regression coefficients from a statistical model and range from 0% to 100%[4]. They are easily derived and intuitive to understand when communicating findings. The calculation of relative importance measures is further described in the Methods section.

Now that attribute importance can be compared on a common scale using relative importance measures, we can rank them to see what attributes are more preferred than others. For example, if attribute A has an importance of 40%, it is twice as important as attribute B with a 20% relative importance, giving it a higher rank. These relative importance measures and rankings, however, may differ depending on the type of model used to analyse the multilevel DCE data [5].

Traditional regression models, that do not adjust for correlated responses, assume individual observations are independent of each other. In DCEs, statistical models such as the multinomial logit model, create three barriers for investigators. They are unable to (i) account for clustered responses within or between individuals; (ii) measure clustering related to observed and unobserved characteristics; and (iii) relax the 'independence of irrelevant alternatives assumption,' which assumes there is no correlation of alternatives across choices[6]. Vass and colleagues identify different approaches to target each of these challenges. For the first challenge, a random effects model can be used; for the second, observed characteristics can be included as covariates; and for the third, a nested model may be more appropriate[6]. A more complex multilevel model such as a latent class or mixed multinomial model can also be used to target all or some of these three challenges.

Understanding the impact of clustering effects on relative importance measures and the ranking of attributes is critical for ensuring that the interpretation of DCE results is accurate, particularly since a variety of adjusted and unadjusted models are currently being used. In some health areas, multilevel models are the leading approach [7-9] while in others they are not[6]. Only a handful of studies use latent class models to adjust for unobserved clusters in the data [6, 8, 10, 11]. Models that do not adjust for observed or unobserved clusters within a heterogeneous dataset may fail to capture correlated responses within and between participants. This may lead to biased conclusions on the attributes' utility estimates and corresponding relative importance measures[12]. An empirical study explored the impact of correlated data on the rank of relative importance measures using several models, however they did not investigate the impact of unobserved heterogeneity via latent class analysis[5]. Their multilevel models were also limited to

only two levels - the choice tasks (level 1) and participants (level 2) [5]. In another study comparing latent class models with mixed logit models, latent class models performed better than mixed logit models; however conclusive statements on which approach was completely superior could not be made [13]. Given the wide array of possible models for making inferences from DCE data, it is important to investigate the robustness of the ranking of relative importance measures in complex multilevel models.

We aimed to explore the sensitivity of the ranking of attributes, and levels within attributes, when different levels of a hierarchical data structure are incorporated in a model. We compared the relative importance rankings of attributes from nine regression models, including fixed effects conditional logit, latent class, and hierarchical latent class models, to explore four different levels of potential cluster variables.

## 2. Methods

Overview

To investigate the impact of several potential cluster variables on the ranking of attributes and attribute levels, we created a fixed effects conditional logit model, and two- and three-level random effects conditional logit models. We describe our methods by first introducing our empirical DCE dataset (anti-bullying program). We then describe part-worth utility estimates, relative importance measures, and ranking of attributes. This is followed by an introduction to latent class analyses. Finally, we outline each model explored and describe how to calculate the relative importance measures of attributes within them.

2.1 Anti-bullying program DCE study (empirical dataset):

Bullying has been tied to developmental risks for both the perpetrators and victims [14, 15]. Cunningham and colleagues set out to explore students' perception of various factors that would influence their peers to intervene when bullying occurs. They conducted a survey of students in elementary schools that was comprised of 15 choice tasks, where each task presented three alternatives. The attributes within each choice task describe characteristics that may affect students' reporting of bullying. Eleven three-level attributes were explored: anonymity of reports, inclusion of students who are left out, interest level of anti-bullying presentations, rewards for preventing bullying, number of playground supervisors, mandatory vs. discretionary reporting, prompt response to reports, reporting by peers, frequency of anti-bullying activities, skill vs empathic content of anti-bullying activities, consequences for perpetrators. Details of each attribute and corresponding three attribute levels are presented in Table 1. Two of the fifteen choice tasks were 'hold out' choice tasks (i.e. the same survey question) to assess the reliability of responses, and thus were excluded in this study's analyses.

The different levels that exist within this dataset's multilevel structure include participants' responses to choice tasks (level 1), individual participants (level 2), classrooms (level 3), grades (level 4), and schools (level 5). Cunningham and colleagues developed a random effects conditional logit model, with six latent classes at the observation level (level 1), two latent group classes at the school level (level 2), and grade as a covariate [16]. This random effects conditional logit model produces part-worth utility estimates for each latent class segment, incorporating participants and schools as random effects. While heterogeneity was incorporated in the statistical

model, the impact of the various group level factors on the relative importance measures of attributes was not investigated.

2.2 Part-worth utility estimates, relative importance measures, and ranking

The use of part-worth utility estimates and relative importance measures in conjoint analysis has previously been described by Orme[4]. We will briefly outline his description below for its relevance to this study.

In DCEs, part-worth estimates within each attribute are scaled to an arbitrary additive constant. For this study, the attribute level part-worth utility estimates were scaled to sum to zero using effects coding. This allows us to compare the value of each level *within* an attribute[4]. A negative utility value does not indicate that an attribute level is unfavourable, it only implies that it is less valued than another level with a positive utility value[4]. Also, due to the arbitrary origin within each attribute, the part-worth utility estimates do not allow us to directly compare values *between* attributes[3].

To compare *between* the 11 attributes, we characterized their relative importance. This is done by considering the difference each attribute potentially has in the total utility of a product, or in this case an anti-bullying program[4]. This difference is the range of the maximum value and the minimum value of levels within an attribute. A percentage from these relative ranges is then calculated to determine a set of relative importance values for each attribute. An attribute's importance is always relative to the other attributes used in the study. We can compare one attribute to another in terms of importance within a study, however, we cannot compare across studies featuring different attribute lists[4].

To compare attributes between different studies and different models, we focus on the ranking of attributes and attribute levels. By ordering the attributes' relative importance measures, we can observe what is most important – or in this case, what will have a larger impact on elementary school students to speak up, report bullying, or seek help from an adult in response to bullying – in comparison to other attributes. We use these relative importance measures to observe how sensitive these rankings are to different models or studies.

## 2.3 Latent class and hierarchical latent class analysis

Latent class analysis is often applied to DCE data to analyze unobserved correlations within responses. Similar responses are partitioned into meaningful classes, where the number of classes and their components are determined during the analysis[17]. Latent class models are comprised of two key components, the class and attribute variables. The class variable indicates the latent or *unobserved* classes of individuals in a population. The attributes are *observed* variables. The latent class model relies on the local independence assumption, where the attributes are mutually independent within each latent class. This implies that the latent variable is the only reason for the correlations - a serious limitation of latent class analyses, since local dependence or correlated responses within classes may also exist[17]. For discrete choice experiments, where choice tasks are completed by each individual, this can lead to similar responses within the data. Ignoring local dependence can lead to spurious latent classes, poor model fit, and reduction in the accuracy of classification [17, 18].

There are several approaches to incorporate local dependence such as with multiple latent variables. Multilevel latent class models contain a hierarchy of latent variables, where group-level

clusters are incorporated[19]. In Cunningham et al's study, several potential layers of clusters exist within their data structure[16]. We used multilevel latent class models to explore their impacts on the final ranking of the relative importance of attributes.

2.4 Outline of regression models

We used several models to explore five different levels of nested data: choice tasks (level 1), participants (level 2), classrooms (level 3), grades (level 4), and schools (level 5). Despite our data potentially having a 5-level hierarchical structure, we were only able to create 2-level and 3-level models within our modelling software, Latent Gold Choice (Version 5.1)[20]. Also, while grade was considered a potential cluster variable, it only had four levels (grades 5,6,7 and 8) and could not be incorporated as a group level latent class variable in our modelling software. Instead, it was explored as a covariate. We created several 2- and 3-level conditional logit models to explore all possible combinations of the five different levels in the data structure. In Latent Gold Choice software, we specified individuals as an individual-level latent variable, and classroom and school as group-level latent class variables. All models contained all 11 three-level attributes with effects coding.

A total of nine models were created, including one fixed effects conditional logit model, and eight random effects conditional logit models. Among the random effects models, four were two-level models and four were three-level models. Models included latent classes at the individual level (level 2), latent classes at a group-level (classroom or school; level 3), or both. Participants' grade was also explored in three models as a covariate. This is further described in Table 2.

The base model within Latent GOLD Choice software is a 2- level model, with responses (level 1) and individuals (level 2). For all models, we were required to specify the number of latent segments at the individual level. If we wanted to ignore individuals as a potential cluster variable, we specified it as having only one-segment. Below we describe how we created each model in the same order as presented in Table 1.

To create *a fixed effects model*, we specified one segment at the individual level in the software. This forced all individuals to be grouped under one class and allowed the model to assume that no heterogeneity existed within the data. This model does not incorporate any of the potential clusters within the data.

To create *a six-segment individual-level latent class model*, we specified six segments at the individual level. The number of segments (6 segments) was predetermined from Cunningham *et al*'s analyses[16]. This model incorporates two levels of the data structure - responses and individuals. We further explored the model with the addition of grade as a covariate.

To create a *one-segment individual-level latent class model with classroom and with school (separately)*, we specified one segment at the individual level, and explored two- and three-segments for classroom and school. These models incorporate two levels of the data structure - responses and classrooms or responses and schools. It was not possible to model classrooms and schools together as group-level latent variables. The best fit model (between the two- and three-segment group level latent variables) was indicated by a lower Bayesian information criterion

(BIC)[21]. Based on the best fit model, this number of segments for the group variables was used in the remaining model.

To create *a six-segment individual-level latent class model with classroom and with school as group level latent variables*, we specified six segments at the individual level and added classroom and school (in separate models) as a group-level latent class variable. These models incorporate three levels of the data structure: responses, individuals, and group (classrooms or schools). We further explored these models with the addition of grade as a covariate.

2.5 Determining the relative importance of attributes for each model

Each model produced mean part-worth utility estimates for each attribute level. For latent class models with 6-segment solutions, six different sets of part-worth utility estimates were estimated, one for each class of individuals. Their mean utility estimates were produced by Latent GOLD by averaging the utility estimates across all six segments. These mean utility estimates were then transformed into relative importance measures to determine the final ranking of attributes for each model.

To calculate the relative importance measures, the range (i.e. the difference between the maximum and minimum utility measure) was first determined for each attribute. This range is then divided by the sum of the ranges of all attributes and multiplied by 100%. Mathematically, for each attribute *k*, the relative importance *(RI)* is determined using the level of attribute *k* with the maximum value $(B_{k,max})$, level with the minimum value$(B_{k,min})$, and the total of the absolute range across all attributes:

$$RI_k = \frac{|\beta_{k,max} - \beta_{k,min}|}{\sum_{j=1}^{11} |\beta_{j,max} - \beta_{j,min}|} \; x \; 100 \;\; [4, 22].$$

Each model's attributes are presented by their corresponding mean part-worth utility estimate and standard deviation (across latent class segments), and relative importance (%). For Models 1, 4, and 5 (Table 2), where only one-segment was specified at the individual-level, the standard error was reported. Comparisons across models were made using the ranking of the attributes' relative importance measures. Attribute levels were ranked using the magnitude of each attribute level's mean part-worth utility measure. This ranking was also compared across models.

In addition to assessing the ranking of attributes from each model's mean part-worth utility estimates, the ranking of latent segments from each latent class model was also assessed (not presented). The attributes' relative importance measures for each latent segment were derived from the part-worth utility measures for each segment using the same above formula.

## 3. Results

Demographics and study details

A total of 2033 students from four different grades (5, 6, 7, and 8), 116 classrooms, and 30 different elementary schools in Ontario, Canada were included in this study. All participants completed 13 choice tasks. Of the participants 479 (23.6%) in grade 5, 536 (26.4%) in grade 6, 533 (26.2%) in grade 7, and 482 (23.7%) in grade 8. Three students reported they were in grade 4 as part of a joint grade 4 and 5 class, who were analyzed as grade 5 students. Females were 1015 (49.9%) of the respondents, males were 803 (39.5%), and 215 (10.6%) preferred not to answer. Students answered various questions exploring how often they witnessed bullying, were victims

of bullying, participated in bullying others, and participated in anti-bullying activities in their schools.

Summary of models

Nine models were created, one fixed effects conditional logit model and eight random effects conditional logit models. The attributes were statistically significant across all models ($p<0.05$). The mean part-worth utility estimates and corresponding standard error or standard deviation for each attribute are presented in Tables 1a and 1b (Appendix). Standard errors are presented for models where no 6-segment latent class analysis at the individual level was conducted. Standard deviations are presented for each 6-segment latent class model and were derived using the part-worth utility estimates of each segment. For models with group-level latent class variables, classroom and school, two segment group-level latent class (2GC) models ($BIC_{2GC\text{-}classroom}$= 48400, $BIC_{2GC\text{-}school}$= 48423) had a better fit than three segment group-level latent class (3GC) models ($BIC_{3GC\text{-}classroom}$= 48511, $BIC_{3GC\text{-}school}$= 48511). Details of each model's goodness-of-fit are presented in Table 3.

Four key points from our analyses

First, the top five anti-bullying program attributes were: prompt response to reports (attribute 7), inclusion of students who are left out (attribute 2), interest level of anti-bullying presentations (attribute 3), reporting by peers (attribute 8), and mandatory vs. discretionary reporting (attribute 6).

Second, the overall ranking of attributes' relative importance measures (based on the mean part-worth utility estimates) remained relatively consistent across models. The relative importance

measures of attributes between models mostly differed by a few decimal places. This resulted in slight differences in their rankings. While attribute 7 remained consistent as the top rank across all nine models, attributes 2, 3, and 8 varied within ranks 2, 3, and 4. Attribute 6 (mandatory vs. discretionary reporting) was consistent as the 5$^{th}$ rank across all models. Details of these rankings with their corresponding relative importance measures are presented in Table 4. The relative importance measures of all 11 attributes are also presented in Figure 1.

Third, the average ranking of levels within each attribute (based on the mean part-worth utility estimates) were robust across all models. That is to say, students consistently chose one level over the other two for each attribute. On average, students were more likely to intervene when: only teachers and the principal know who reported the bullying (attribute 1), students always included left out students (attribute 2), anti-bullying activities are interesting (attribute 3), students who try to prevent bullying are rewarded (attribute 4), four teachers watch the playground (attribute 5), students are asked to report bullying (attribute 6), there is an immediate school response after a student reports bullying (attribute 7), students in the school almost always report bullying (attribute 8), anti-bullying activities are once a month (attribute 9), students are taught how to stop bullying (attribute 10), and bullies get suspended for a week (attribute 11).

Fourth, the ranking of the part-worth utility measures from each latent class segment differed from the ranking of their mean part-worth utility, across all latent classes. Of the six latent segments within each latent class model, the attributes to consistently appear as top rank in at least one latent segment were (i) inclusion of students who are left out and (ii) consequences for perpetrators. The top ranked attributes in other latent segments, less consistently were: mandatory vs. discretionary reporting, reporting by peers, frequency of anti-bullying activities, and interest

level of anti-bullying presentations – some of which were included in the top five attributes. The models do identify similar groups of students in the latent class models. The ranking within each 6 latent segment is similar across each model, especially for models that only differ by the inclusion of grade as a covariate.

## 4. Discussion

Overview

This study set out to investigate the impact of clustering effects on the ranking of relative importance measures of 11 three-level attributes in a DCE study empirical dataset. The relative importance measures were derived from the mean part-worth utility estimates of nine regression models: one fixed regression model and eight random effects conditional logit models. Some of the random effects models adjusted for clustering using latent class segments at the individual level, at a group level (classroom or school), and at both the individual and group level. The impact of grade was also investigated as a covariate. Overall, the rankings of the relative importance measures, based on mean part-worth utility estimates, across all nine models were relatively robust. The rankings of attribute levels within each attribute, based on mean part-worth utility estimates, were identical across models. For models that adjusted for six latent classes at the individual level, similarities were seen between the six segments (or six 'types of individuals') of each model, however, the attribute rankings between segments in each model differed.

Methodological studies within the health literature of DCEs compare models either with the ranking of relative importance measures or focus on comparing models using other approaches. Andrews and colleagues conducted a simulation study to compare finite mixture models and

hierarchical Bayesian estimation models to explore discrete versus continuous representations of heterogeneity. While they found that both were equally effective in determining individual level parameter estimates and predicting ratings of hold-out tasks, the models' impact on relative importance measures were not explored [23]. Similarly, Greene and Hensher compared a multinomial logit model with a mixed multinomial logit model and a 3-class latent class model on a DCE dataset [24]. While the mixed logit model was a better fit model than the multinomial logit model in terms of the log likelihood values, it was difficult to compare the mixed model with the latent class model. No conclusive results were made except that they supported both models for incorporating unobserved heterogeneity. Cheng and colleagues investigated the impact of nine different models: six types of logit and probit models for a binary outcome, one bivariate probit model for two correlated binary outcomes, and three multinomial logit and probit models for a nominal outcome [5]. They found a similar ranking of relative importance measures between models when the DCE data had a low clustering effect, determined using the intraclass correlation coefficient (ICC; where ICC≈0). More variation in the rankings between models was observed when the data had a higher clustering effect (ICC=0.659). Our findings complement Cheng and colleagues' empirical comparison of various multinomial logit and probit models[5].

This study has several strengths. We further explore the sensitivity of students' choice of schools where they were more likely to speak up, report bullying, or seek help from an adult (in response to bullying). This was done using a comprehensive analysis of nine DCE models, including a conditional logit model, latent class models, and hierarchical latent class models. Several layers of heterogeneity were accounted for from responses nested within individuals and individuals nested within classrooms, grades, and schools. To our knowledge, this is also the first

empirical study to investigate the impact of multilevel latent class models on the ranking of attributes.

A few limitations arise in this study. First, grade was added as a covariate and not a group-level latent variable. Latent GOLD software requires a larger number of levels for group-level latent variables, and since grade only had 4 levels (5,6,7 and 8), it was more appropriate to include it as a covariate. Second, for the group-level latent variables, we were limited to only exploring 2- and 3-segment solutions. Third, while we identified a potential 5-level data structure (responses, participants, classrooms, grades, and schools), we were only able to create 3-level models. Fourth, the focus of this study was to empirically compare the impact of various hierarchical models on the ranking of relative importance measures - we did not explore other covariates (outside of grade) or the interactions between attributes or attribute levels.

Further research could be conducted to validate the findings from this study. First, it would be valuable to conduct a simulation study to explore how robust the ranking of attributes are when various degrees of correlated responses exist within the data. Second, adjusting for potential confounders at the observation level, such as participant demographics, should also be investigated. Third, exploring a similar study in different types of DCEs, such as best-worst designs[25], full or fractional factorial designs[26], with various characteristics of the design, such as number of participant, attributes, and attribute levels, would reveal how sensitive the rankings are in different DCE design settings.

## 5. Conclusions

This study explored the robustness of the ranking of attributes' relative importance measures across various hierarchical models, from an empirical dataset. Using the mean part-worth utility estimates to derive these relative importance measures, the ranking of attributes were similar and the ranking of levels within attributes were consistent across models. For 6-segment latent class models, differences were observed in the rankings between individual latent classes and their overall ranking across latent classes. A simulation study to explore how robust the final rankings of attributes are in various degrees of correlated responses would provide further insight into the impact of heterogeneity on the analysis and interpretation of DCEs.

## 6. Tables and Figures

Table 1: List of attributes and attribute levels

| Attribute | Attribute level |
|---|---|
| 1 Anonymity of reports | Only teachers and the principal know who reports bullying |
| | Only the principal knows who reports bullying |
| | No one knows who reports bullying |
| 2 Inclusion of students who are left out | Students never include left out students |
| | Students sometimes include left out students |
| | Students always include left out students |
| 3 Interest level of anti-bullying presentations | Anti-bullying activities are boring |
| | Anti-bullying activities are okay |
| | Anti-bullying activities are interesting |
| 4 Rewards for preventing bullying | Students who try to prevent bullying are not rewarded |
| | Students who try to prevent bullying are sometimes rewarded |
| | Students who try to prevent bullying are rewarded |
| 5 Number of playground supervisors | 2 teachers watch the playground |
| | 4 teachers watch the playground |
| | 8 teachers watch the playground |
| 6 Mandatory vs. discretionary reporting | Asks students to report bullying |
| | Tells students they have to report bullying |
| | If students don't report bullying they get in trouble |
| 7 Prompt response to reports | When students report bullying, this school responds immediately |
| | When student report bullying, this school responds the next day |
| | When students report bullying, this school responds in one week |
| 8 Reporting by peers | Students almost never report bullying |
| | Students sometimes report bullying |
| | Students almost always report bullying |
| 9 Frequency of anti-bullying activities | Anti-bullying activities are every day |
| | Anti-bullying activities are once a month |
| | Anti-bullying activities are twice a year |
| 10 Skill vs empathic content of AB activities | Tells students "don't bully" |
| | Teaches students how bullying affects victims |
| | Teaches students how to stop bullying |
| 11 Consequences for perpetrators | Teachers just talk to bullies |
| | Bullies lose recess for a week |
| | Bullies get suspended for a week |

Table 2: Breakdown of conditional logit regression models

| Conditional logit models | Levels of structure | Random effect(s) considered | # Individual-level Latent Segments | # Group-level Latent Segments | Covariate |
|---|---|---|---|---|---|
| 1. Fixed effects | 1 | None | 1 | N/A | N/A |
| 2. Six-segment ILC | 2 | ID | 6 | N/A | N/A |
| 3. Six -segment ILC with grade (covariate) | 2 | ID | 6 | N/A | Grade |
| 4. One-segment ILC with classroom (GLC) | 2 | Classroom | 1 | 2, 3 | N/A |
| 5. One-segment ILC with school (GLC) | 2 | School | 1 | 2, 3 | N/A |
| 6. Six -segment ILC with classroom (GLC) | 3 | ID, Classroom | 6 | 2, 3 | N/A |
| 7. Six -segment ILC with classroom (GLC) and grade (covariate) | 3 | ID, Classroom | 6 | 2, 3 | Grade |
| 8. Six -segment ILC with school (GLC) | 3 | ID, School | 6 | 2, 3 | N/A |
| 9. Six -segment ILC with school (GLC) and grade (covariate) | 3 | ID, School | 6 | 2, 3 | Grade |

*ILC=Individual-level latent class; GLC=group-level latent class; ID=individual

Table 3: Model fit for each model explored

| Model | Random effects (±grade as covariate) | LL | BIC | Npar |
|---|---|---|---|---|
| 1 | None | | | |
| | *1-Class Choice* | -24111.5 | 48390.6 | 22 |
| 2 | Responses, Individual | | | |
| | *6-Class Choice* | -22755.7 | 46555.0 | 137 |
| 3 | Responses, Individual (+Grade) | | | |
| | *6-Class Choice* | -22728.2 | 46614.2 | 152 |
| 4 | Classroom | | | |
| | *1-Class 2-GLC Choice* | -24028.6 | 48400.0 | 45 |
| | *1-Class* 3-GLC Choice | -23996.3 | 48510.5 | 68 |
| 5 | School | | | |
| | *1-Class 2-GLC Choice* | -24040.2 | 48423.1 | 45 |
| | *1-Class 3-GLC Choice* | -23996.3 | 48510.5 | 68 |
| 6 | Individual, Classroom | | | |
| | *6-Class 2-GLC Choice* | -22683.1 | 46623.1 | 165 |
| 7 | Individual, Classroom (+Grade) | | | |
| | *6-Class 2-GLC Choice* | -22658.6 | 46688.3 | 180 |
| 8 | Individual, School | | | |
| | *6-Class 2-GLC Choice* | -22677.2 | 46611.3 | 165 |
| 9 | Individual, School (+Grade) | | | |
| | *6-Class 2-GLC Choice* | -22649.0 | 46669.1 | 180 |

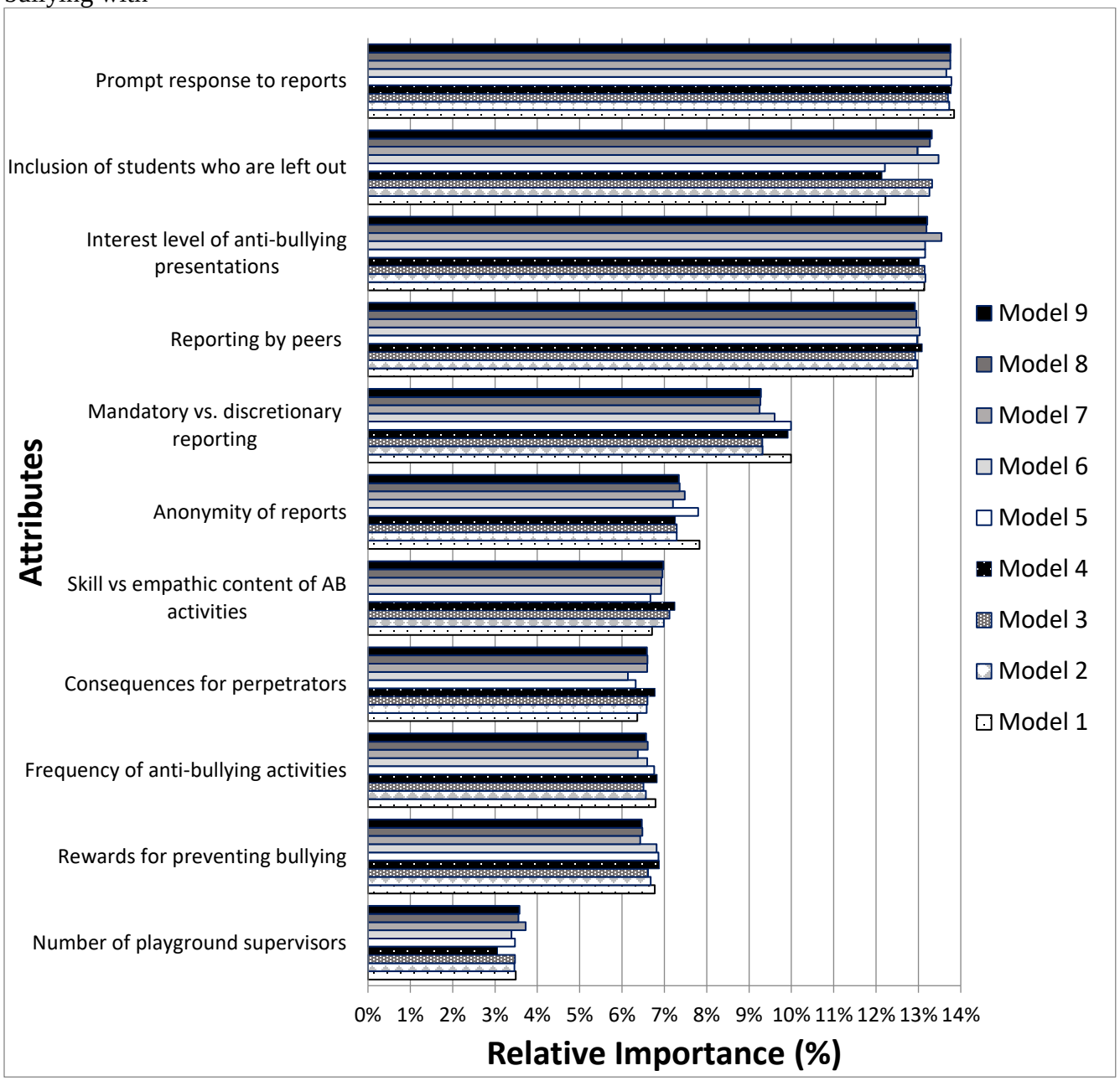
Comments: LL=log-likelihood; BIC=Bayesian information criteria; Npar=number of parameters; 1-Class models are the default models of Latent Gold the entire dataset is considered as one class; 6-Class models are for latent classes at the individual choice task response level and GLC indicates latent classes at a higher group level (for classroom and school). Group level variables were first investigated with 2 and 3 GLCes in Models 4 and 5. After deciding that a 2GLC models had a better fit, these were used in Models 6-9. Grade was included as a covariate.

Table 4: Order of attributes' relative importance measures from highest to lowest for all models

| Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 | Model 8 | Model 9 |
|---|---|---|---|---|---|---|---|---|
| **Fixed effects** | **6-segment ILC** | **6-segment ILC with grade (covariate)** | **1-segment ILC with classroom (GLC)** | **1-segment ILC with school (GLC)** | **6-segment ILC with classroom (GLC)** | **6-segment ILC with classroom (GLC) and grade (covariate)** | **6-segment ILC with school (GLC)** | **6-segment ILC with school and grade (covariate)** |
| **Attribute # (RI)** | **Attribute # (RI)** | **Attribute # (RI)** | **Attribute # (RI)** | **Attribute # (RI)** | **Attribute # (RI)** | **Attribute # (RI)** | **Attribute # (RI)** | **Attribute # (RI)** |
| 7(13.8) | 7(13.7) | 7(13.7) | 7(13.8) | 7(13.8) | 7(13.7) | 7(13.8) | 7(13.8) | 7(13.8) |
| 3(13.1) | 2(13.3) | 2(13.3) | 8(13.1) | 3(13.2) | 2(13.5) | 3(13.5) | 2(13.3) | 2(13.3) |
| 8(12.9) | 3(13.2) | 3(13.1) | 3(13.0) | 8(13.0) | 3(13.2) | 2(13.0) | 3(13.2) | 3(13.2) |
| 2(12.2) | 8(13.0) | 8(12.9) | 2(12.1) | 2(12.2) | 8(13.0) | 8(13.0) | 8(13.0) | 8(12.9) |
| 6(10.0) | 6(9.3) | 6(9.3) | 6(9.9) | 6(10.0) | 6(9.6) | 6(9.2) | 6(9.3) | 6(9.3) |
| 1(7.8) | 1(7.3) | 1(7.3) | 1(7.3) | 1(7.8) | 1(7.2) | 1(7.5) | 1(7.4) | 1(7.3) |
| 9(6.8) | 10(7.0) | 10(7.1) | 10(7.2) | 4(6.9) | 10(6.9) | 10(6.9) | 10(7.0) | 10(7.0) |
| 4(6.8) | 4(6.7) | 4(6.6) | 4(6.9) | 9(6.8) | 4(6.8) | 11(6.6) | 9(6.6) | 11(6.6) |
| 10(6.7) | 11(6.6) | 11(6.6) | 9(6.8) | 10(6.7) | 9(6.6) | 4(6.4) | 11(6.6) | 9(6.6) |
| 11(6.4) | 9(6.6) | 9(6.5) | 11(6.8) | 11(6.3) | 11(6.1) | 9(6.4) | 4(6.5) | 4(6.5) |
| 5(3.5) | 5(3.5) | 5(3.5) | 5(3.1) | 5(3.5) | 5(3.4) | 5(3.7) | 5(3.6) | 5(3.6) |

Comment: ILC=latent class model; GLC=group level latent classes; ID=individual; RI= relative importance; Results are reported as attribute number and corresponding relative importance measure (%) in brackets. Each attribute by number corresponds to: 1=Anonymity of reports;2= Inclusion of students who are left out; 3=Interest level of anti-bullying presentations; 4=Rewards for preventing bullying; 5=Number of playground supervisors; 6=Mandatory vs. discretionary reporting; 7=Prompt response to reports; 8 Reporting by peers; 9=Frequency of anti-bullying activities; 10=Skill vs empathic content of AB activities; 11=Consequences for perpetrators.

Figure 1: Relative importance measures of all eleven attributes students are most likely to intervene bullying with




**Comment: Model 1:** Fixed effects conditional logit; **Model 2**: Six-segment ILC; **Model 3:** Six-segment ILC with grade (covariate); **Model** 4: One-segment ILC with classroom (GLC); **Model 5:** One-segment ILC with school (GLC); **Model 6:** Six-segment ILC with classroom (GLC); **Model 7:** Six-segment ILC with classroom (GLC) and grade (covariate); **Model 8:** Six-segment ILC with school (GLC); **Model 9:** Six-segment ILC with school (GLC) and grade (covariate).

## References

1. Madhulatha TS: **An overview on clustering methods**. *arXiv preprint arXiv:12051117* 2012.
2. Lancsar E, Louviere J: **Conducting Discrete Choice Experiments to Inform Health Care Decision Making: A User's Guide, 2008**. *Pharmacoeconomics* 2008, **26**.
3. Lancsar E, Louviere JJ, Flynn TN: **Several methods to investigate relative attribute impact in stated preference experiments**. *Social Science and Medicine* 2007, **64**.
4. Orme B: **Interpreting the Results of Conjoint Analysis**. In: *Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research, Second Edition.* edn. Madison, Wis: Research Publishers, LLC; 2010: 77-89.
5. Cheng J, Pullenayegum E, Marshall DA, Marshall JK, Thabane L: **An empirical comparison of methods for analyzing correlated data from a discrete choice survey to elicit patient preference for colorectal cancer screening**. *BMC medical research methodology* 2012, **12**(1):15.
6. Vass C, Gray E, Payne K: **Discrete choice experiments of pharmacy services: a systematic review**. *International journal of clinical pharmacy* 2016, **38**(3):620-630.
7. Ryan M, Gerard K: **Using discrete choice experiments to value health care programmes: current practice and future research reflections**. *Applied health economics and health policy* 2003, **2**(1):55-64.
8. de Bekker-Grob EW, Ryan M, Gerard K: **Discrete choice experiments in health economics: a review of the literature**. *Health Econ* 2012, **21**(2):145-172.
9. Mandeville KL, Lagarde M, Hanson K: **The use of discrete choice experiments to inform health workforce policy: a systematic review**. *BMC Health Services Research* 2014, **14**(1):367.
10. 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.
11. Zhou M, Thayer WM, Bridges JF: **Using Latent Class Analysis to Model Preference Heterogeneity in Health: A Systematic Review**. *PharmacoEconomics* 2017:1-13.
12. Goossens LMA, Utens CMA, Smeenk FWJM, Donkers B, van Schayck OCP, Rutten-van Mölken MPMH: **Should I Stay or Should I Go Home? A Latent Class Analysis of a Discrete Choice Experiment on Hospital-At-Home**. *Value in Health* 2014, **17**(5):588-596.
13. Shen J: **Latent class model or mixed logit model? A comparison by transport mode choice data**. *Applied Economics* 2009, **41**(22):2915-2924.
14. Arseneault L, Walsh E, Trzesniewski K, Newcombe R, Caspi A, Moffitt TE: **Bullying victimization uniquely contributes to adjustment problems in young children: a nationally representative cohort study**. *Pediatrics* 2006, **118**(1):130-138.
15. Kim YS, Leventhal BL, Koh YJ, Hubbard A, Boyce WT: **School bullying and youth violence: causes or consequences of psychopathologic behavior?** *Archives of general psychiatry* 2006, **63**(9):1035-1041.
16. Cunningham CE, Rimas H, Thabane L: **What Program Designs Motivate Student Intervention in Bullying Episodes: Multilevel Latent Class Analysis of a Discrete Choice Conjoint Experiment in Grade 5, 6, 7, and 8**. In.
17. Zhang NL: **Hierarchical latent class models for cluster analysis**. *Journal of Machine Learning Research* 2004, **5**(6):697-723.
18. Vermunt JK, Magidson J: **Latent class cluster analysis**. *Applied latent class analysis* 2002, **11**:89-106.
19. Vermunt JK: **Latent class and finite mixture models for multilevel data sets**. *Statistical Methods in Medical Research* 2008, **17**(1):33-51.

20. Vermunt JK, Magidson J: **Technical Guide for Latent GOLD 5.1: Basic, Advanced, and Syntax**. In. Belmont, MA: Statistical Innovations Inc; 2016.
21. Hauber AB, González JM, Groothuis-Oudshoorn CG, Prior T, Marshall DA, Cunningham C, IJzerman MJ, Bridges JF: **Statistical methods for the analysis of discrete choice experiments: a report of the ISPOR Conjoint Analysis Good Research Practices Task Force**. *Value in Health* 2016, **19**(4):300-315.
22. Sullivan LM, Massaro JM, D'Agostino RB: **Presentation of multivariate data for clinical use: the Framingham Study risk score functions**. *Statistics in medicine* 2004, **23**.
23. Andrews RL, Ainslie A, Currim IS: **An empirical comparison of logit choice models with discrete versus continuous representations of heterogeneity**. *Journal of Marketing Research* 2002, **39**(4):479-487.
24. Greene WH, Hensher DA: **A latent class model for discrete choice analysis: contrasts with mixed logit**. *Transportation Research Part B: Methodological* 2003, **37**(8):681-698.
25. Lancsar E, Louviere J, Donaldson C, Currie G, Burgess L: **Best worst discrete choice experiments in health: methods and an application**. *Social science & medicine (1982)* 2013, **76**(1):74-82.
26. Jaynes J, Wong WK, Xu H: **Using Blocked Fractional Factorial Designs to Construct Discrete Choice Experiments for Health Care Studies**. *Statistics in medicine* 2016, **35**(15):2543-2560.

# APPENDIX

Table 1a: Mean part-worth utility estimates of attributes for each model

| Attribute | | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|---|
| | | Mean(SE) | Mean(SD) | Mean(SD) | Mean(SE) | Mean(SE) |
| **1 Anonymity of reports** | Only teachers and the principal know who reports bullying | 0.48(0.02) | 0.58(0.26) | 0.57(0.24) | 0.46(0.02) | 0.48(0.02) |
| | Only the principal knows who reports bullying | 0.02(0.02) | 0.02(0.11) | 0.02(0.09) | 0.04(0.02) | 0.02(0.02) |
| | No one knows who reports bullyinzg | -0.50(0.02) | -0.59(0.24) | -0.6(0.24) | -0.49(0.03) | -0.5(0.03) |
| **2 Inclusion of students who are left out** | Never include | -0.74(0.03) | -1.06(0.70) | -1.06(0.71) | -0.76(0.03) | -0.74(0.03) |
| | Sometimes include | -0.06(0.02) | -0.02(0.22) | -0.01(0.21) | -0.06(0.03) | -0.06(0.03) |
| | Always include | 0.80(0.02) | 1.07(0.76) | 1.08(0.76) | 0.82(0.02) | 0.80(0.02) |
| **3 Interest level of anti-bullying presentations** | Boring | -0.94(0.03) | -1.21(0.58) | -1.21(0.57) | -0.98(0.04) | -0.95(0.04) |
| | Okay | 0.24(0.02) | 0.31(0.2) | 0.31(0.21) | 0.26(0.03) | 0.24(0.03) |
| | Interesting | 0.71(0.02) | 0.90(0.47) | 0.9(0.48) | 0.72(0.02) | 0.71(0.02) |
| **4 Rewards for preventing bullying** | Not rewarded | -0.47(0.02) | -0.60(0.41) | -0.59(0.38) | -0.49(0.03) | -0.48(0.03) |
| | Sometimes rewarded | 0.08(0.02) | 0.13(0.18) | 0.12(0.18) | 0.08(0.02) | 0.09(0.02) |
| | Always rewarded | 0.38(0.02) | 0.47(0.36) | 0.47(0.34) | 0.41(0.02) | 0.39(0.02) |
| **5 Number of playground supervisors** | 2 teachers watch the playground | -0.26(0.02) | -0.33(0.30) | -0.33(0.30) | -0.22(0.03) | -0.25(0.03) |
| | 4 teachers watch the playground | 0.18(0.02) | 0.22(0.12) | 0.23(0.10) | 0.18(0.02) | 0.19(0.02) |
| | 8 teachers watch the playground | 0.07(0.02) | 0.11(0.32) | 0.1(0.32) | 0.04(0.03) | 0.07(0.03) |
| **6 Mandatory vs. discretionary reporting** | Asks students to report bullying | 0.45(0.02) | 0.53(0.33) | 0.53(0.32) | 0.46(0.02) | 0.45(0.02) |
| | Tells students they have to report bullying | 0.36(0.02) | 0.43(0.18) | 0.43(0.18) | 0.36(0.02) | 0.36(0.02) |
| | If students don't report bullying they get in trouble | -0.81(0.03) | -0.97(0.39) | -0.96(0.39) | -0.83(0.03) | -0.81(0.03) |
| **7 Prompt response to reports** | School responds immediately | 0.87(0.02) | 1.08(0.51) | 1.08(0.49) | 0.89(0.02) | 0.87(0.02) |
| | School responds the next day | <0.01(0.02) | 0.05(0.18) | 0.04(0.17) | 0.03(0.03) | <0.01(0.03) |
| | School responds in one week | -0.87(0.03) | -1.12(0.57) | -1.12(0.57) | -0.91(0.04) | -0.87(0.04) |
| **8 Reporting by peers** | Almost never | -0.83(0.03) | -1.09(0.63) | -1.08(0.62) | -0.88(0.04) | -0.85(0.04) |
| | Sometimes | 0.05(0.02) | 0.09(0.23) | 0.09(0.23) | 0.06(0.03) | 0.05(0.03) |
| | Almost always | 0.78(0.02) | 0.99(0.69) | 0.99(0.69) | 0.83(0.03) | 0.79(0.03) |
| **9 Frequency of anti-bullying activities** | Every day | -0.01(0.02) | -0.09(0.82) | -0.09(0.81) | -0.09(0.03) | -0.02(0.03) |
| | Once a month | 0.43(0.02) | 0.57(0.44) | 0.57(0.45) | 0.49(0.03) | 0.44(0.03) |
| | Twice a year | -0.42(0.02) | -0.48(0.52) | -0.48(0.49) | -0.4(0.03) | -0.42(0.03) |
| **10 Skill vs empathic content of AB activities** | Tells students "don't bully" | -0.54(0.03) | -0.72(0.50) | -0.74(0.53) | -0.61(0.03) | -0.54(0.03) |
| | Teaches students how bullying affects victims | 0.24(0.02) | 0.33(0.26) | 0.33(0.29) | 0.27(0.02) | 0.24(0.02) |
| | Teaches students how to stop bullying | 0.30(0.02) | 0.40(0.26) | 0.41(0.27) | 0.34(0.02) | 0.3(0.02) |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Teachers just talk to bullies | -0.44(0.02) | -0.61(0.83) | -0.61(0.81) | -0.47(0.03) | -0.44(0.03) |
| **11 Consequences for perpetrators** | Bullies lose recess for a week | 0.07(0.02) | 0.16(0.57) | 0.17(0.55) | 0.06(0.02) | 0.08(0.02) |
| | Bullies get suspended for a week | 0.36(0.02) | 0.45(0.99) | 0.45(0.97) | 0.41(0.02) | 0.36(0.02) |
| **BIC (based on LL)** | | 48390.6 | 46555.0 | 46614.1 | 48400.0 | 48423.1 |
| **AIC (based on LL)** | | 48267.0 | 45785.4 | 45760.3 | 48147.2 | 48170.3 |
| **AIC3 (based on LL)** | | 48289.0 | 45922.4 | 45912.3 | 48192.2 | 48215.3 |
| **CAIC (based on LL)** | | 48412.6 | 46692 | 46766.1 | 48445.0 | 48468.1 |

Comment: **Model 1:** Fixed effects conditional logit; **Model 2**: Six-segment LCM; **Model 3:** Six-segment LCM with grade (covariate); **Model 4**: One-segment LCM with classroom (GCLASS); **Model 5:** One-segment LCM with school (GCLASS); **Model 6:** Six-segment LCM with classroom (GCLASS); **Model 7:** Six-segment LCM with classroom (GCLASS) and grade (covariate); **Model 8:** Six-segment LCM with school (GCLASS); **Model 9:** Six-segment LCM with school (GCLASS) and grade (covariate); CL=conditional logit; LCM=latent class model; GCLASS=group level latent classes; SD= standard deviation; SE=standard error; All p-values for each attribute were <0.001 indicating that they were statistically significant. Complete descriptions of attribute levels are presented in Table 1 of the manuscript

Table 1b: Mean part-worth utility estimates of attributes and pvalues for each model (continued)

| | Attribute | Model 6 | Model 7 | Model 8 | Model 9 |
|---|---|---|---|---|---|
| | | Mean(SD) | Mean(SD) | Mean(SD) | Mean(SD) |
| **1 Anonymity of reports** | Only teachers and the principal know who reports bullying | 0.57(0.28) | 0.59(0.24) | 0.59(0.26) | 0.59(0.27) |
| | Only the principal knows who reports bullying | 0.01(0.15) | 0.02(0.09) | 0.02(0.10) | 0.02(0.11) |
| | No one knows who reports bullyinzg | -0.59(0.24) | -0.61(0.22) | -0.6(0.25) | -0.60(0.24) |
| **2 Inclusion of students who are left out** | Never include | -1.08(0.73) | -1.04(0.73) | -1.07(0.72) | -1.07(0.72) |
| | Sometimes include | -0.02(0.24) | -0.01(0.24) | -0.02(0.26) | -0.01(0.24) |
| | Always include | 1.10(0.79) | 1.05(0.78) | 1.08(0.79) | 1.08(0.79) |
| **3 Interest level of anti-bullying presentations** | Boring | -1.22(0.58) | -1.25(0.61) | -1.22(0.58) | -1.22(0.59) |
| | Okay | 0.31(0.22) | 0.31(0.21) | 0.30(0.19) | 0.31(0.20) |
| | Interesting | 0.91(0.47) | 0.93(0.51) | 0.91(0.49) | 0.91(0.49) |
| **4 Rewards for preventing bullying** | Not rewarded | -0.62(0.43) | -0.58(0.38) | -0.59(0.38) | -0.58(0.38) |
| | Sometimes rewarded | 0.14(0.17) | 0.12(0.19) | 0.12(0.17) | 0.12(0.17) |
| | Always rewarded | 0.48(0.36) | 0.46(0.32) | 0.46(0.34) | 0.46(0.33) |
| **5 Number of playground supervisors** | 2 teachers watch the playground | -0.32(0.33) | -0.37(0.29) | -0.35(0.26) | -0.35(0.27) |
| | 4 teachers watch the playground | 0.23(0.13) | 0.23(0.11) | 0.23(0.10) | 0.23(0.11) |
| | 8 teachers watch the playground | 0.09(0.34) | 0.13(0.31) | 0.12(0.30) | 0.12(0.29) |
| **6 Mandatory vs. discretionary reporting** | Asks students to report bullying | 0.55(0.33) | 0.53(0.30) | 0.53(0.32) | 0.53(0.32) |
| | Tells students they have to report bullying | 0.44(0.18) | 0.43(0.19) | 0.44(0.19) | 0.44(0.19) |
| | If students don't report bullying they get in trouble | -1.00(0.39) | -0.96(0.39) | -0.97(0.40) | -0.97(0.40) |
| **7 Prompt response to reports** | School responds immediately | 1.08(0.51) | 1.09(0.52) | 1.10(0.53) | 1.09(0.51) |
| | School responds the next day | 0.05(0.21) | 0.03(0.18) | 0.03(0.19) | 0.03(0.17) |
| | School responds in one week | -1.13(0.58) | -1.12(0.58) | -1.13(0.59) | -1.13(0.59) |
| **8 Reporting by peers** | Almost never | -1.10(0.65) | -1.09(0.65) | -1.09(0.65) | -1.09(0.64) |
| | Sometimes | 0.10(0.24) | 0.09(0.23) | 0.09(0.23) | 0.09(0.22) |
| | Almost always | 1.00(0.71) | 1.00(0.71) | 1.0(0.71) | 1.00(0.71) |
| **9 Frequency of anti-bullying activities** | Every day | -0.11(0.83) | -0.05(0.83) | -0.06(0.80) | -0.06(0.81) |
| | Once a month | 0.59(0.45) | 0.54(0.41) | 0.56(0.40) | 0.56(0.41) |
| | Twice a year | -0.48(0.50) | -0.49(0.53) | -0.51(0.53) | -0.50(0.52) |
| **10 Skill vs empathic content of AB activities** | Tells students "don't bully" | -0.72(0.50) | -0.72(0.53) | -0.73(0.52) | -0.73(0.53) |
| | Teaches students how bullying affects victims | 0.33(0.27) | 0.31(0.28) | 0.33(0.26) | 0.33(0.28) |
| | Teaches students how to stop bullying | 0.40(0.25) | 0.40(0.27) | 0.4(0.27) | 0.40(0.27) |
| **11 Consequences for perpetrators** | Teachers just talk to bullies | -0.58(0.84) | -0.61(0.81) | -0.61(0.83) | -0.61(0.82) |
| | Bullies lose recess for a week | 0.18(0.57) | 0.15(0.58) | 0.15(0.55) | 0.15(0.55) |
| | Bullies get suspended for a week | 0.41(0.99) | 0.45(1.00) | 0.46(1.01) | 0.46(1.01) |
| **BIC (based on LL)** | | 46623.1 | 46688.3 | 46611.3 | 46669.1 |
| **AIC (based on LL)** | | 45696.3 | 45677.2 | 45684.5 | 45658 |
| **AIC3 (based on LL)** | | 45861.3 | 45857.2 | 45849.5 | 45838 |
| **CAIC (based on LL)** | | 46788.1 | 46868.3 | 46776.3 | 46849.1 |

Comment: **Model 6:** 6-segment LCM with classroom (GCLASS), **Model 7:** 6-segment LCM with classroom (GCLASS) and grade (covariate), **Model 8:** 6-segment LCM with school (GCLASS), **Model 9:** 6-segment LCM with school and grade (covariate); CL=conditional logit; LCM=latent class model; GCLASS=group level latent

classes; SD= standard deviation; SE=standard error; All p-values for each attribute were <0.001 indicating that they were statistically significant. Complete descriptions of attribute levels are presented in Table 1 of the manuscript.

# CHAPTER 6

# SOME DISCUSSIONS AND CONCLUSIONS

This section summarises the findings from the thesis by addressing the research questions that guided this work. We also highlight implications for research, practice and policy; and comment on the substantive contributions of this thesis.

## Part 1: Addressing the research questions

### 1. What do DCE simulation studies conclude on the impact of various design features on relative design efficiency?

We found nine simulation studies that reported relative design efficiency in our systematic survey. Design features that improved statistical efficiency of a DCE were (i) increasing the number of choice tasks or alternatives; (ii) decreasing the number of attributes, and levels within attributes; (iii) using model-based designs with covariates, or sampling approaches that incorporate response behaviour; (iv) incorporating heterogeneity in a model-based design; (v) correctly specifying Bayesian priors and minimizing parameter prior variances; and (vi) using an appropriate method to create the DCE design for the research question[1].

### 2. What is the quality of reporting of DCE simulation studies?

The nine studies identified in our systematic survey were of good quality, however further clarifications and specifications are needed. All nine studies identified in our systematic survey clearly reported the primary outcome, rationale, methods to create the designs, and methods to

evaluate each scenario. The total number of designs and the range of design characteristics explored were either written or easily identifiable from figures and tables. Most studies reported the software used for the simulation studies[1].

Some information was lacking within the literature. Reporting of research objectives was unclear in two studies. There was limited to no reporting of any failures in the simulations, the starting seeds of their simulations, the number of times they were unable to create a design given the design component restrictions. Improved reporting of detailed study objectives, random number generators, starting seeds, number of failures, and the software used is needed.

**3. How will varying the number of choice tasks, alternatives, attributes, and levels within attributes impact the statistical efficiency of a DCE?**

We created and measured the statistical efficiency of 3204 DCE designs, using relative design efficiency. We found that relative design efficiency increased as the number of attributes and attribute levels decreased, and as the number of alternatives increased. Relative design efficiency converged as the number of choice tasks increased, but not necessarily to 100% statistical optimality. Further, if the number of alternatives was a multiple of the number of choice tasks in a design, higher values of relative design efficiency - often 100% - were achieved[2].

**4. How robust is the ranking of attributes when differed fixed effects and random effect approaches are used to analyse DCE data — namely, multinomial logit (MNL) and multinomial probit (MNP) models?**

We assessed the sensitivity of participant preferences on 16 knowledge translation variables, using empirical data. Attributes' relative importance measures were derived from fixed effects MNL and MNP models, and random effects MNL and MNP models. Through an empirical comparison analysis, we found that the ranking of attributes, by relative importance, is minimally influenced by models adjusting for clustering effects or correlations in participants' responses to choice tasks. Ranking remained relatively robust, regardless of the statistical model used to analyse the data. Differences in ranking between fixed effects and random effects models often occurred due to very small differences in relative importance measures. This, however, may be due to low correlations within this specific empirical dataset. In addition, there are potential unobserved correlations within the data that may impact the ranking of attributes, which were not explored in this study.

**5. How robust is the ranking of attributes when various approaches are used adjust for multi-level clustering are used — namely fixed effects conditional logit model, latent class models, and hierarchical latent class models?**

We assessed the sensitivity of elementary students' choice of anti-bullying programs in schools, described by 11 attributes with three levels each, using empirical data. We created nine models, one fixed effects conditional logit model, and eight random effects conditional logit models incorporating latent classes. Each latent class model had six segments at the individual

level. The sensitivity of the attributes' rankings were based on their relative importance measures. These relative importance measures were derived in two ways: (i) using each model's mean part-worth utility estimates, and (ii) using the part-worth utility estimates from each of the six latent classes within each model. Finally, we explored the ranking of the levels within attributes, using their mean part-worth utility estimates.

The overall ranking of the 11 attributes' relative importance measures, based on their mean part-worth utility estimates, were similar across all nine models. Where differences in ranking were observed, this was often caused by differences in relative importance measures by a few decimal places.

For latent class models, attributes' rankings based on their part-worth utility estimates (for each segment) differed from the ranking derived from their mean part-worth utility estimates (across all segments). This was consistent across all models.

The average ranking of the three levels within each attribute, based on their mean part-worth utility estimates, were robust across all models.

In summary, the rankings of attributes' relative importance measures were robust across several models adjusting for various layers of heterogeneity. In exploring potential unobserved clusters within the data, we found that they impact the ranking of attributes. Thus, variations in rankings occur within these models adjusting for latent classes.

**Part 2: Implications for research, practice, and policy regarding the design and analysis of DCEs**

**Implications for research:**

Our research identified several features that impact statistical efficiency, measured by design efficiency, which can be used to design future DCEs. Using the knowledge of these design patterns and relevant characteristics of the population of interest, researchers can design more optimal DCEs that maximize both statistical efficiency and response efficiency.

Further studies can be done to explore optimal DCE designs. First, a systematic review of simulation studies investigating all measures of statistical efficiency will ensure that these findings are robust across different measures. Second, comparisons of optimal designs generated by different approaches will assess the sensitivity of how design features impact statistical efficiency. Third, further incorporating both statistical efficiency and response efficiency in simulation studies will develop more pragmatic designs. Fourth, there needs to be more discussion of blocked DCE designs, which may improve statistical efficiency.

We explored rankings in attributes' relative importance measures - across a variety of models, using two empirical datasets - and found differences that should be further investigated by researchers when analyzing DCEs. When using mean part worth utility estimates, we observed similar rankings, but when using part worth utility estimates from the individual segments of latent class models, the rankings differed. The similarities found using the mean part worth utility estimates, however, may have been be due to low correlations within the data. Further research can be done to explore sensitivity of attributes' rankings using different methods to analyze DCEs.

There are several ways this can be done, including (i) conducting simulation studies and additional analyses on empirical datasets, particularly with various degrees of correlated responses within the data; (ii) adjusting for potential confounders, such as participant demographics; and (iii) conducting empirical comparisons with different DCE designs, such as best-worst designs, full factorial designs, and fractional factorial designs, as well as with different design features, such as number of participant, attributes, and attribute levels.

Finally, our systematic survey of nine simulation studies highlights critical criteria that are seldom reported, which researchers need to improve on.

**Implications for practice and policy:**

Health care providers and policy makers - who use DCEs to make inferences about patient and public preferences regarding health services, products, or programs - can use our research findings to better evaluate the applicability of DCEs. Our findings suggest that the rankings of attributes are fairly consistent despite using different models to analyse the data. Thus, we can make inferences from several DCEs, even if they use different models, to come up with the most preferred attributes across a population. This, however, is limited to the average of the populations. Not all studies adjust for clustering within their models, particularly latent clustering. Our research reveals differences in preferences among latent class groups. If one is planning to make health care decisions or policies relevant to a subset of a population or to a highly heterogeneous population, it may be important to use a DCE that incorporates latent class analysis.

**Some concluding remarks:**

In this thesis, we review and summarize the literature of DCE simulation studies, identify limitations within the reporting of the literature, explore numerous designs to evaluate their impact on statistical efficiency, identify patterns within the design features, and evaluate the sensitivity of the rankings of relative importance measures using several models. Throughout this thesis, we employed a number of study designs including a systematic survey, simulation study, and methodological comparisons on empirical data, to create a body of evidence to compliment the current DCE literature. We hope that this thesis serves are as an aid for future research in the development, application, analysis of DCEs.

## References

1. Vanniyasingam T, Daly C, Jin X, Zhang Y, Foster G, Cunningham C, Thabane L: **Investigating the impact of design characteristics on statistical efficiency within discrete choice experiments: A systematic survey**. *Contemporary clinical trials communications* 2018, **10**:17-28.
2. Vanniyasingam T, Cunningham CE, Foster G, Thabane L: **Simulation study to determine the impact of different design features on design efficiency in discrete choice experiments**. *BMJ open* 2016, **6**(7):e011985.