Conditional, Structural and Unobserved Heterogeneity: three essays on preference heterogeneity in the design of financial incentives to increase weight loss program reach

Yuan Yuan

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Economics, Agriculture and Life Sciences

Kevin J Boyle, Co-Chair
Wen You, Co-Chair
Paul A Estabrooks
Christopher F Parmeter

May 27, 2015
Blacksburg, Virginia

Keywords: discrete choice, preference heterogeneity, attribute non-attendance, mixed logit

Conditional, Structural and Unobserved Heterogeneity: three essays on preference heterogeneity in the design of financial incentives to increase weight loss program reach

Yuan Yuan

(ABSTRACT)

This dissertation consists of three essays on forms of preference heterogeneity in discrete choice models.

The first essay uses a model of heterogeneity conditional on observed individual-specific characteristics to tailor financial incentives to enhance weight loss program participation among target demographics. Financial incentives in weight loss programs have received attention mostly with respect to effectiveness rather than participation and representativeness. This essay examines the impact of financial incentives on participation with respect to populations vulnerable to obesity and understudied in the weight loss literature. We found significant heterogeneity across target sub-populations and suggest a strategy of offering multiple incentive designs to counter the dispersive effects of preference heterogeneity.

The second essay investigates the ability of a novel elicitation format to reveal decision strategy heterogeneity. Attribute non-attendance, the behaviour of ignoring some attributes when performing a choice task, violates fundamental assumptions of the random utility model. However, self-reported attendance behaviour on dichotomous attendance scales has been shown to be unreliable. In this essay, we assess the ability of a polytomous attendance scale to ameliorate self-report unreliability. We find that the lowest point on the attendance scale corresponds best to non-attendance, attendance scales need be no longer than two or three points, and that the polytomous attendance scale had limited success in producing theoretically consistent results.

The third essay explores available approaches to model different features of unobserved heterogeneity. Unobserved heterogeneity is popularly modelled using the mixed logit model, so called because it is a mixture of standard conditional logit models. Although the mixed logit model can, in theory, approximate any random utility model with an appropriate mixing distribution, there is little guidance on how to select such a distribution. This essay contributes to suggestions on distribution selection by describing the heterogeneity features which can be captured by established parametric mixing distributions and more recently introduced nonparametric mixing distributions, both of a discrete and continuous nature. We provide empirical illustrations of each feature in turn using simple mixing distributions which focus on the feature at hand.

This work is supported by the Virginia Tech Institute for Critical Technology and Applied

Science (ICTAS) and the National Institutes of Health (NIH-NHLBI R21, PI: You).

Dedication

To my father, who led me onto this lifelong path of learning, and has always been there for me.

Acknowledgments

I would like to thank my advisors, Dr Kevin Boyle and Dr Wen You, for their guidance and support during my PhD. They were true mentors to me, showing me how to conduct research, navigate academia, and prepare for my future. Everything I know of being a true scholar, I owe to them. At the same time, their unceasing care and concern as human beings nurtured me on this path. Where I would have faltered, they bolstered me. Without them I would not have been able to complete this journey.

I would also like to thank my committee members. The perspective of Dr Paul Estabrooks was invaluable to keeping my work grounded and applicable, while the experience and advice of Dr Christopher Parmeter was invaluable to my technical efforts. The coding discussions we shared were some of the most enjoyable aspects of my work.

I would like to thank the Virginia Tech Institute for Critical Technology and Applied Science (ICTAS) for their generous financial support, enabling me to pursue my PhD.

I would like to thank my colleagues. Sapna Kaul, Ranju Baral and Kristen Capogrossi for being inspiring role models, Xiaoshu Li, Shuwen Duan and Jackie Yenerall for being compatriots with whom I could share the highs and lows, and especially Quentin Stoeffler and Jianfeng Gao for going through everything together, as well as Nadezda Amaya and Aditya Khanal for forging the camaraderie that brought me such happiness. Special thanks to Fiona Achi for being my 'study buddy'.

Finally I would like to thank my family. My parents, Yew Yuan and Ning Zhao, who have given me so much. My dog, Donut, who provided necessary companionship, stress relief and an excuse to exercise and be outside every day. My partner, Hidekazu Miwa, for supporting me day in and day out, in ways small and large, from the start to the finish. I can imagine no greater proof of his love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Heterogeneity is a fundamental aspect of preferences. It is self-evident; stepping outside of academia and into daily life, the thought of finding two individuals who are identical in their preferences seems almost ludicrous. While accounting for preference heterogeneity may not be needed in all cases, it is often useful to equity considerations in policymaking and necessary to avoid biased estimates. Through three essays, this dissertation explores different forms of preference heterogeneity and suggests approaches for analyzing each form.

All three essays use as a case study a choice experiment on the design of financial incentives accompanying a behavioural weight loss program. Choice experiments are a type of stated preference valuation method which elicit preferences over goods or services in hypothetical markets. They present hypothetical goods or services which are defined by a set of attributes and the levels those attributes take, a presentation consistent with Lancaster's theory of characteristics of goods (Lancaster, 1966). Respondents to the choice experiment choose their preferred good or service as if they were in a market situation, thus yielding choice data which can be analyzed with discrete choice models.

Discrete choice models operationalize preferences via the random utility model, which assumes that individuals choose alternatives from a choice set to maximize utility (McFadden, 1974). Utility may be affected by factors as diverse as subjective perception of the attribute levels, previous experience with the good or service, assumptions extrapolating characteristics of the good or service beyond the presented attribute levels, and private information like situational constraints. Since the analyst is unable to perfectly observe all the factors which affect choice, utility is endowed with a stochastic component. Consequently, choice models identify choice only up to a probability.

Although heterogeneity in the stochastic component (*e.g.*, scale heterogeneity) has been studied, many forms of heterogeneity affect the systematic component of the random utility, including the ones considered in this dissertation. The first essay uses a model of heterogeneity conditional on observed individual-specific characteristics to tailor financial incentives to enhance weight loss program participation among target demographics. The second essay investigates the ability of a novel elicitation format to reveal decision strategy heterogeneity. The third essay explores available approaches to model different features of unobserved heterogeneity.

The first essay is motivated by the policy goal of enhancing *reach* (Glasgow et al., 1999) among individuals who are vulnerable to overweight and obesity or who have been understudied in the weight loss literature. Incorporating financial incentives into behavioural weight loss programs have been previously suggested as a strategy to increase the efficacy of the programs, but few studies have considered the impact of financial incentives on the participation rates and representativeness of participants of the programs. Even fewer have specifically examined the impact of financial incentives differentially, with respect to populations vulnerable to obesity or understudied in the weight loss literature. Discrete choice models with heterogeneity conditional on individual-specific characteristics are ideally suited to emphasize both participation and representativeness, since choice probabilities can be interpreted as probability of participation and demographic-specific participation probabilities can be predicted.

Using a conditional logit model with interactions between socioeconomic covariates and alternative-specific variables, we found important policy implications of preference heterogeneity across the target demographic groups. Although reward amount and program location were the most important attributes for all demographic groups, there was consistent evidence that, relatively speaking, the African-American sub-population preferred non- or minimally-monetized incentives. At the same time, the obese sub-population were very responsive to the reward amount. Our results revealed a trade-off in participation from those who respond positively and those who respond negatively to reward amount, indicating a low reward amount may balance the two groups, attracting the former while minimizing alienation of the latter. However, a better strategy may be to offer multiple incentive designs, which increases the likelihood that one of the designs will attract an individual enough to participate in the weight loss program. Although this strategy has little effect on those who are responsive to the monetary component of the incentive, such as the obese sub-population, it does no harm and greatly increases the participation rates among others.

The second essay is motivated by the unreliability of self-reported attribute attendance behaviour. Attribute non-attendance is the behaviour of ignoring some attribute when performing a choice task in a choice experiment. This behaviour may present a serious problem for modelling and inference, because it violates fundamental assumptions of the random utility model. However, self-reported attendance has been found to be inconsistent with theoretical expectations, perhaps due to misreporting, conflation between low and zero preferences, and variable attendance behaviour across choice tasks. Polytomous attendance scales, which allow respondents to select a degree of attribute attendance rather than forcing a binary yes/no response, have the potential to address these sources of unreliability.

Using conditional logit models in which the coefficients on ignored attributes were constrained to zero, we found that the lowest point or lowest two points on the polytomous attendance scale were the best indicators of non-attendance. Such a specification assumes that self-reported attendance behaviour is reliable, and to assess this reliability, we used conditional logit models with interactions between self-reported attendance level and alternative-specific variables. We found that a six-point polytomous attendance scale contained no more information than a shorter two- or three-point attendance scale, and further, that none of the scales (polytomous, dichotomous, or trichotomous) displayed consistency between empirical observations and theoretical expectations.

The third essay is motivated by the plethora of probability distributions available to describe unobserved preference heterogeneity. Unobserved heterogeneity is often captured via random coefficients in the systematic component of random utility. The choice model derived under these conditions is known as the mixed logit model, because it can be seen as a mixture of choice probability expressions under the standard conditional logit model. The mixing distribution for the mixed logit model arises from the analyst-specified choice of distribution for the random coefficients, and in the past decade, the number and complexity of proposed mixing distributions has increased greatly. Within this large array of mixing distributions, we recommend analysts choose based on the heterogeneity features which are relevant according to policy and theoretical perspectives.

We describe established parametric mixing distributions and more recently introduced nonparametric mixing distributions, with an emphasis on properties which capture different heterogeneity features, such as boundedness, modality, symmetry and dependence between random coefficients. We further divide the nonparametric mixing distributions into continuous and discrete distributions, suggesting that the choice between continuous and discrete representations of heterogeneity is now much more nuanced, depending on the policy context, research question and intended audience. Using the choice experiment on financial incentives in weight loss programs, we empirically illustrate the search for each heterogeneity feature. We found strong evidence for dependence between random coefficients, not only of the linear, correlation type, but also more complex dependence structures. Our evidence for other features was weaker, in part due to estimation difficulties with some of the mixing distributions.

These three essays are tied by a common theme of preference heterogeneity, but each considers a different form of heterogeneity and fulfills a different purpose. Although preference heterogeneity is a fact, the forms we impose on it are not real – they are artificial, constructed to shed light on specific research questions. In the first essay, we use heterogeneity conditional on observed demographics to target specific sub-populations with customized programs. In the second essay, we attempt to improve the detection of decision strategy heterogeneity, to avoid confounding taste variation with other forms of heterogeneity. In the third essay, we provide a guide to the many approaches for describing unobserved heterogeneity, which, after other forms of heterogeneity have been accounted for, captures taste variation.

Bibliography

Glasgow, RE, Vogt, TM, and Boles, SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *Am J Public Health*, 89(9):1322–1327, September 1999. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1508772/.

Lancaster, Kelvin J. A new approach to consumer theory. *Journal of Political Economy*, 74 (2):pp. 132–157, 1966. ISSN 00223808. URL http://www.jstor.org/stable/1828835.

McFadden, Daniel. Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in econometrics*, chapter 4, pages 104–142. Academic Press, 1974.

Chapter 2

Designing financial incentives to enhance participation of target populations in weight loss programs

Abstract

Incorporating financial incentives into behavioural weight loss programs have been previously suggested as a strategy to increase the efficacy of the programs, but few studies have considered the impact of financial incentives on the participation rates and representativeness of participants of the programs. Even fewer have specifically examined the impact of financial incentives differentially, with respect to populations vulnerable to obesity and understudied in the weight loss literature. This study uses a discrete choice experiment to investigate preferences among target subgroups for the design of financial incentives in a behavioural weight loss program. We found the presence of significant heterogeneity, with the African-American subpopulation more interested in non-monetary incentives than other groups. In contrast, we found the obese sub-population highly responsive to the monetary amount of the incentive. We suggest a strategy of offering multiple incentive designs to counter the dispersive effects of preference heterogeneity.

2.1 Introduction

The rapid and persistent rise in obesity prevalence across the nation (and the world) is widely considered a public health crisis. Recent projections based on US data indicate that all American adults will become overweight or obese by 2048 if current trends continue, and that the healthcare costs attributable to overweight and obesity will reach almost one trillion dollars in 2030, nearly one-fifth of total US healthcare costs (Wang et al., 2008). The need for broad-based and effective weight loss interventions has never been greater.

The WHO suggests that an important component of behavioural weight loss programs is non-food-based rewards for achievements in weight loss (WHO Consultation on Obesity, 2000). The rationale for using rewards arises from both operant conditioning in psychology (Jeffery, 2012) and behavioural economics in economics (Cawley and Price, 2011). Weight loss is a complex, long-term phenomenon, and the path from weight loss behaviour to weight loss itself is not always straightforward. Operant conditioning and behavioural economics explain how rewards may help bridge the gap between weight loss behaviour and weight loss achievement by providing immediacy, saliency, consistency and commitment. Consequently, the use of financial incentives as a form of reward in behavioural weight loss programs has compelled the interest of researchers for decades (*e.g.*, Mann, 1972; Forster et al., 1985; Jeffery and French, 1997; Volpp et al., 2008; Kullgren et al., 2013).

The vast majority of the studies on financial incentives in weight loss programs have focused on efficacy in convenience samples, even though "recruiting at-risk groups and individuals is the first step in an effective weight-management protocol" (WHO Consultation on Obesity, 2000). RE-AIM, a framework to evaluate and enhance translating research into practice, formalizes this idea in its first dimension, "reach" (Glasgow et al., 1999). Reach is composed of two components: participation and representativeness.

Participation refers to the absolute number as well as the proportion of willing participants, and is important as a prerequisite for a high impact public health intervention. No matter how efficacious the intervention is, it cannot have a high impact on public health without willing participants. Representativeness refers to the how well the willing participants represent the target population, *e.g.*, at-risk populations for overweight and obesity, or those already overweight or obese. Representativeness is an important factor in improving health and healthcare equity. Research which ignores representativeness will serve those most willing to be studied or most easily studied, not those who most need the intervention.

Participation is particularly relevant in the context of obesity, because the high prevalence of overweight and obesity calls for high impact programs. High impact programs require high rates of participation, but only a handful of studies among dozens on financial incentives in weight loss programs considered the effect of incentives on participation rate. Jeffery et al. (1990) was the only study to have found a significant effect of treatment condition on participation rate. In their study, two treatment conditions were offered, paying a $5 program fee or placing a $60 deposit which would be refunded upon success in the program. The authors found that although the deposit contract induced twice as much weight loss, the fee contract induced five times as much enrolment. This finding illustrates the potential for tension between targeting participation vs targeting effectiveness. In this particular case, the results indicate that adopting the approach which induces higher enrolment would ultimately have a greater impact than adopting the approach which induces greater effectiveness (multiplying the community population, recruitment rate and average weight loss implies 5,950 lb weight lost in the community under the fee contract vs 912 lb under the deposit contract).

Cawley and Price (2011) and Butsch et al. (2007) investigated the impact on participation of a control condition versus an incentive condition (deposit contract in the former study, partial reimbursement of program fees in the latter study). In contrast to Jeffery et al. (1990), neither study found a significant difference in enrolment rates between the treatment conditions. On the other hand, Jeffery et al. (1983) and Forster et al. (1985) found results similar to Jeffery et al. (1990), in that increased burden was associated with lower enrolment rates. Jeffery et al. (1983) offered six treatment conditions varying along two dimensions, deposit amount and reward conditions. They found that a lower deposit amount was associated with a higher enrolment rate. Forster et al. (1985) offered four treatment conditions varying along two dimensions, group vs self instruction and required vs optional attendance at weigh-ins and sessions. They found that requiring attendance was associated with lower enrolment rates in males. However, these two studies had relatively small sample sizes and the results were not significant. Taken together, these studies suggest the possibility that incentives which may be perceived as being more onerous may lower participation rates, although the findings are not conclusive.

The other component of reach, representativeness, is also particularly relevant in the context of obesity because overweight and obesity affect some groups more than others. In the US, African Americans and low-income women have disproportionately high rates of overweight and obesity (Flegal et al., 2012; Ogden et al., 2010). However, participants in weight loss studies are disproportionately female (Pagoto et al., 2012), Caucasian, highly educated and have higher incomes (Sherwood et al., 1998). Participants in studies on financial incentives in weight loss programs follow suit, with a few exceptions. Jeffery and French (1997, 1999), specifically recruited low-income women in addition to other participants, and Relton et al. (2011) oversampled residents of socioeconomically deprived areas. Jeffery and French (1997) did not find significant treatment effects by participant type (*i.e.*, men, high-income women and low-income women), although they comment on some insignificant trends seen in weight gain. Men and high-income women showed less weight gain with more intervention, but low-income women unexpectedly showed more weight gain with more intervention. Jeffery and French (1997) reported on the intermediate progress of a 3-year intervention after 1 year, whereas Jeffery and French (1999) reported on the same intervention when it was complete after 3 years. However, Jeffery and French (1999) did not report on treatment effect by participant type, so whether the insignificant trends found in Jeffery and French (1997) persisted to program completion is unknown. Relton et al. (2011) also did not report treatment effect by social deprivation score. Consequently, the results cannot be used to target vulnerable populations, even though the study sample demonstrated representativeness. In order to close health disparities, we must not only include health-disadvantaged populations in our studies, but also consider treatment effect interactions. If, for example, low-income women react differently to interventions, as Jeffery and French (1997) suggest is a possibility, we need to implement interventions which are tailored to this group, rather than the population average, in order to pursue health equity.

We can only speculate as to why participation and representativeness are understudied in the literature on financial incentives in weight loss. All of the studies in the literature focus primarily on treatment effectiveness, and consequently, participation is only a secondary consideration if it is considered at all. Evaluating participation in the context of treatment effectiveness is difficult because oftentimes, information about the sampling frame or population of interest is insufficient. Many studies in this literature used relatively small sample sizes, and hence suffered from weak power in statistical tests on interaction effects between treatment and demographics. Larger studies, on the other hand, were sometimes conducted in the workplace context and so the samples were relatively homogeneous. Ensuring representativeness often requires oversampling, since some sub-populations tend to display lower response rates or be excluded from studies due to criteria such as illiteracy and co-morbidities. Consequently, what is needed to inform participation and representativeness in the area of financial incentives in weight loss is a study which focuses on participation rather than effectiveness, provides sufficient power to detect interaction effects between demographic characteristics and intervention options, and uses a clearly defined sampling frame which oversamples demographics of interest.

Discrete choice experiments answer those needs, and are thus well-suited to studying both participation and representativeness. Discrete choice experiments are a method for economic valuation, used particularly when actual market behaviour cannot be observed (because, *e.g.*, the market does not exist). They elicit preferences for a hypothetical good or service (*e.g.*, financial incentives in weight loss programs) by presenting versions of that good or service which vary in its characteristics, and asking survey respondents to choose which version they prefer. In choice experiment language, each good or service is defined by its attributes (*e.g.*, payment type) and each version of the good or service takes on specific level of that attribute (*e.g.*, debit card or gym pass). The attribute levels are varied systematically according to an experimental design. Lancsar and Louviere (2008) identify discrete choice experiments as a valuable tool in health economics for evaluating participation and predicting uptake in new policies or products. Thus, discrete choice experiments can be used to understand which characteristics of the financial incentives are desirable, and thus how the design of the incentive itself can increase or decrease participation rates.

Furthermore, a much richer exploration of incentive preferences is possible with discrete choice experiments than with traditional studies focused on treatment effectiveness rather than treatment preferences. Previous studies on financial incentives in weight loss programs each used only a very limited set of treatment conditions. The majority of studies included only a single treatment condition (*i.e.*, weight loss program with financial incentive) compared to a control condition (*i.e.*, weight loss program *without* financial incentive), equivalent to studying a single factor with a single level (Mahoney, 1974; Jeffery et al., 1990; Cameron et al., 1990; Jeffery et al., 1993a; Jeffery and French, 1997, 1999; Jeffery et al., 1998; Heshka et al., 2003; Hubbert et al., 2003; Butsch et al., 2007; John et al., 2011; Relton et al., 2011; Petry et al., 2011; Crane et al., 2012; Lahiri and Faghri, 2012; Moller et al., 2012). Other studies used a single factor, but with two levels (Mann, 1972; Wing et al., 1981; Jeffery et al., 1984; Forster et al., 1985; Jeffery et al., 1985; Kramer et al., 1986; Jeffery et al., 1993b; Volpp et al., 2008; Augurzky et al., 2012; Reuss-Borst et al., 2012; Kullgren et al., 2013) or three levels instead (Jeffery et al., 1978; Finkelstein et al., 2007). A few studies have used two factors with two or three levels for each factor (Saccone and Israel, 1978; Jeffery et al., 1983; Mavis and Stöffelmayr, 1994), and no study has used more than two factors. As far as the author is aware, six (2×3) is the maximum number of treatment conditions in any study design (Jeffery et al., 1983). The 'attributes' in discrete choice experiments are equivalent to the 'factors' in general experiments. In a literature review of discrete choice experiment applications in health economics from 1990–2000, Ryan and Gerard (2003) found that studies used two to twenty-four attributes, with most studies including four to six attributes. Hence, discrete choice experiments typically consider a far greater number of factors and levels than previous studies on financial incentives in weight loss programs.

One consequence of the limited number of factors and levels used in the design of previous studies is the lack of comparability across studies. Treatment conditions varied from study to study, and because each study contained so few treatment conditions, they rarely overlapped with each other. Among those studies which included more than one treatment condition, many factors were studied: reward condition, positive or negative incentive, type of individual giving reward, magnitude, social aspect of treatment, reward schedule, treatment schedule, reward type, and certainty of reward. Hence, few studies included the same factors. Moreover, even if two studies included a common factor, the levels of that factor were often different. For example, reward condition, one of the more popular factors, could take levels such as weight loss (Jeffery et al., 1978; Saccone and Israel, 1978; Forster et al., 1985; Mavis and Stöffelmayr, 1994), attendance at instructional sessions or weigh-ins (Forster et al., 1985; Kramer et al., 1986; Mavis and

Stöffelmayr, 1994), a combination of both weight loss and attendance (Kramer et al., 1986), calorie restriction (Jeffery et al., 1978) and change in eating behaviour (Saccone and Israel, 1978). Thus, very few studies compare treatment across the same factors and levels, and the findings from single treatment condition studies are even more difficult to generalize and compare across studies. The literature suffers from wide variation across studies not only with respect to incentive configurations, but also with respect to the weight loss interventions which the financial incentives accompany (Paul-Ebhohimhen and Avenell, 2008; Burns et al., 2012; Jeffery, 2012). As a result, arriving at conclusive findings is difficult.

Discrete choice experiments are also appropriate for studying the representativeness of participants in weight loss programs with financial incentives. Discrete choice experiments are implemented as surveys, and thus are usually cheaper to implement than interventions. Consequently, larger sample sizes and more factors and levels in the experimental design are easier to attain. Additionally, we can oversample target populations to ensure adequate representation of their preferences in the choice data. Discrete choice theory then provides a simple way of estimating interactions between treatment conditions and individual characteristics. Understanding subgroup preferences allows us to customize health interventions for target populations, encouraging representativeness.

Subgroup treatment effects have rarely been studied. Among the few studies which considered interaction effects between treatment conditions and demographic variables, gender was the most popular individual characteristic studied. Most of those studies found a significant gender interaction effect (Jeffery et al., 1984; Forster et al., 1985; Hubbert et al., 2003; Cawley and Price, 2013; Reuss-Borst et al., 2012), but a few did not (Jeffery et al., 1993b; Finkelstein et al., 2007). Other characteristics whose interaction effects have been studied include age (Jeffery et al., 1985; Hubbert et al., 2003; Finkelstein et al., 2007; Cawley and Price, 2013), job category (Jeffery et al., 1985, 1993b) and race (Finkelstein et al., 2007), but none of these interaction effects were found to be significant. Among previous studies on financial incentives in weight loss programs, one is particularly notable for its extensive consideration of treatment effect heterogeneity. Reuss-Borst et al. (2012) found that increasing the magnitude of the financial reward increased weight loss among women, migrants, those in good health and those who frequently cook at home. Moreover, women, singles, unemployed and those who seldom cook at home were induced to lose weight with financial rewards, but did not do so in the absence of a financial reward.

This study uses a discrete choice experiment to elicit preferences for financial incentive designs in a behavioural weight loss program, conditional on individual characteristics. The incentive designs are defined using five attributes: reward magnitude, program location, payment form, reward condition and payment frequency. The weight loss intervention which the financial incentive accompanies is held fixed across all versions of the financial incentive, and is a standard behavioural weight loss program (Pinto et al., 2007). Overweight or obese adults were recruited for the study, and low socioeconomic status individuals and males were oversampled. Relevant demographic characteristics were interacted with incentive design attributes in order to model subgroup preferences. This study contributes to the literature in several ways: 1) focusing on participation rather than effectiveness, 2) considering a far greater number of factors and levels than previous designs, while holding the weight loss intervention itself fixed, 3) oversampling populations at risk for overweight and obesity or understudied in the previous literature, 4) modelling subgroup preferences so that interventions can be customized for target populations.

2.2 Methods

In this section we describe the methods we use to model and interpret the preferences for financial incentive designs in a behavioural weight loss program. We first begin with an overview of discrete choice theory, which provides the theoretical basis and analysis approach for data generated by discrete choice experiments. We describe how the presence of individual heterogeneity (*e.g.*, interactions between individual characteristics and incentive attributes) can be detected and modelled. We then interpret the discrete choice model by examining the importance of each attribute to different subgroups and the subgroup preferences for attribute levels within each attribute. This interpretation can be used to design incentives which will be attractive to target sub-populations.

2.2.1 Discrete choice theory

Choice experiment data are typically analyzed with choice models (Holmes and Adamowicz, 2003), which can be derived from the random utility model (RUM). RUMs are based on the theory of characteristics of goods (Lancaster, 1966), in which preferences are defined over the attributes of the goods rather than the goods themselves. Specifically, the random utility an individual i receives from by choosing an alternative j is defined as

$$U_{ij} = V_{ij} + \epsilon_{ij}$$

that is, the sum of a deterministic component, the systematic utility V_{ij}, and a stochastic component, the error term ϵ_{ij}.

Choice is modelled by comparing the random utility of different alternatives for the same individual. That is, an individual i is assumed to choose alternative j if the random utility of that alternative U_{ij} exceeds the random utility of all other alternatives in the choice set, $U_{im}, m \neq j$. Hence, the probability that individual i chooses alternative j is $p_{ij} = \Pr(U_{ij} > U_{im}) \forall m \neq j$. The specification of the error term determines the type of choice model. For our primary model, we specify the error term to be iid Gumbel-distributed, resulting in the conditional logit model, which has the following choice probability expression:

$$p_{ij} = \frac{\exp(V_{ij})}{\sum_m \exp(V_{im})}$$

We use a straightforward linear additive specification for the systematic utility in the primary conditional logit model.

$$V_{ij} = X_{ij}\beta$$

where X_{ij} are the independent variables and β are the preference parameters on those variables. X_{ij} includes main effect terms for the alternative-specific terms (variables associated with the attributes), main effect terms for the individual-specific terms (demographic characteristics), and first order interaction terms between the alternative-specific and individual-specific terms. The inclusion of interaction terms allows attribute effects to differ by demographic characteristics.

2.2.2 Preference heterogeneity

To establish the presence of preference heterogeneity, we use a secondary model in the form of a random parameters logit model. The presence of preference heterogeneity suggests that it may be captured by observable characteristics, which can be used to customize incentive designs and inform targeted policies. A random parameters logit model allows preference parameters to differ across individuals, by specifying them as being drawn from a probability distribution, such as the normal distribution. Its choice probability expression is also in the form of a multinomial logit, save that the logit probability is an expectation with respect to the assumed preference parameter distribution:

$$p_{ij} = \int \frac{\exp(V(X_{ij};\beta))}{\sum_m \exp(V(X_{im};\beta))} f(\beta) d\beta$$

where X contain variables representing the attribute levels and β are the preference parameters for those attribute levels. Hence, the secondary random parameters logit model is an attributes-only model which does not include any observed demographic variables, but is used to establish the need for including demographic variables into the primary conditional logit model.

If the presence of preference heterogeneity is established, we can try to capture some of that heterogeneity using interactions with demographic variables in the primary conditional logit model. As discussed in the introduction, previous studies have found that gender and employment status may have significant interactions with treatment conditions. Furthermore, obesity and overweight disproportionately affect African Americans and low-income women, so relevant demographic characteristics associated with target populations are race, gender and income. We also argue for the inclusion of educational attainment on intuitive grounds. We expect that health literacy and numeracy are important in comprehending the survey, the weight loss program presented, and the financial incentives, particularly their attribute levels. Educational attainment should serve as a proxy for health literacy and numeracy. Combining these findings with characteristics of at-risk populations, the relevant demographic variables which are available from our survey instrument are race, gender, income, employment status and education. We also include weight status because the foremost target population of a broad-based weight loss intervention individuals is those who are already overweight and obese. Most subsequent analysis is then conducted using the primary conditional logit model, which captures preference heterogeneity using interaction terms between attributes and the named demographic variables.

2.2.3 Attribute importance

In the context of translational obesity research, attribute importance, in addition to preference heterogeneity, is a useful output of choice models. If an attribute is relatively unimportant, then implementations of the incentive design may either ignore it or set it to a level convenient for implementation (rather than a level which is most preferred by potential participants), thus decreasing implementation cost. Conversely, if an attribute is relatively important, then policy implementations should ensure that it is set to the optimal level to target desired populations.

In this study, we use partial log-likelihoods to determine attribute importance because this approach assesses the impact of entire attributes rather than individual attribute levels (Lancsar et al., 2007). This approach proxies attribute importance with the relative explanatory power of the attribute, as measured by differences between the log-likelihood value of the full model and the log-likelihood value of a reduced model in which the attribute has been removed. Attributes which are more important, in some sense, will contribute more explanatory power to the model leading to larger differences between the full and partial log-likelihood values.

Specifically, we estimate partial models, in which all variables associated with a particular attribute is dropped from the specification. We compute the difference between the log-likelihood of the partial model and the log-likelihood of the full model, which we call the partial effect. These partial effects are computed for every attribute in the choice experiment.

We perform the partial log-likelihood analysis separately for each demographic sub-sample of interest: obese, African Americans, low-income women (females with household income $40,000 or less), and males. These demographic groups are based on the sub-populations which are vulnerable to obesity or understudied in the weight loss literature. Since the related individual-specific covariates are unidentified for the sub-samples (*e.g.*, the BMI status covariate is unidentified for the obese sub-sample, gender is unidentified for the male sub-sample), we apply the partial log-likelihood analysis to a secondary, attributes-only conditional logit model. We do not use the primary conditional logit model in this part of the analysis because the full model specification would necessitate removing the unidentified individual-specific covariates, resulting in a different model specification for each sub-sample. Consequently, the model results would not be comparable across sub-samples.

2.2.4 Attribute level preferences

After ranking the attributes by importance, we also need to determine the relative preference for levels within each attribute, conditional on relevant demographic variables. The relative preferences for attribute levels suggest the optimal combination of attributes and levels for target demographic groups which will maximize the probability of participation. The relative preferences for attribute levels can be determined using probability analysis, in which the participation probabilities under a base incentive design and a design in which one attribute takes on a different level are compared (Lancsar et al., 2007). In the context of understanding the role of financial incentives in weight loss program participation decisions, the most natural base case is the opt-out alternative. Hence, we modify the probability analysis to focus on the participation probability (the probability that respondents choose not to opt out) as a measure of the 'marginal effect' of a given attribute level.

Specifically, we consider a choice set with two alternatives, the opt-out alternative and a design alternative which differs from the opt-out alternative by only one attribute level. Using the primary conditional logit model, we predict the probability that the design alternative will be chosen, and call this probability the participation probability. We systematically define design alternatives for every attribute level, and rank the attribute levels by predicted participation probability within each attribute.

The reward amount attribute deserves special attention here. Not only is it likely to be of considerable policy interest since it could be the primary cost driver in implementation, but it is also the only continuous attribute and thus requires somewhat different analysis. Participation probability could very well monotonically increase with the reward amount, at least for some individuals, thus precluding the existence of an optimum level. Instead, it may be more useful to investigate the responsiveness of participation probability to reward amount, conditional on relevant demographic variables. Some subgroups may show greater response than others, indicating for whom increasing reward amount will have the most effect.

As a point of clarification, note that the attribute level analysis is based entirely on predicted choice probabilities from the primary conditional logit model (that is, with a full model specification including all interactions with demographic variables). Since subgroup analysis conducted on predicted choice probabilities rather than at the level of the estimation dataset, identification is no longer a problem as it was for the attribute importance analysis.

Furthermore, analysis at the level of predicted choice probabilities allows generalizing beyond the study sample. We treat each unique combination of demographic variables as an *individual profile* (*e.g.*, an employed, white female with high education, high income and normal weight), and appropriately weighting each individual profile yields results representative of the sub-population of interest. Thus, the predicted participation probabilities are specific not only to design alternatives but also to individuals. Hence, the rankings of attribute levels within attributes are also individual-specific. Using appropriate weights, these individual-specific rankings can be summarized in a representative way.

2.3 Data

In this section we describe the development of the survey instrument and choice experiment and the implementation of the survey. Attributes and their levels used in the choice experiment were developed through both literature review and focus groups. The choice experiment was conducted in the context of a standard behavioural weight loss intervention. The description of the intervention was identical across all versions of the survey instrument; only the financial incentive design itself was varied across choice tasks and survey instrument versions. The choice experiment was conducted via mail survey, and participants were recruited from obese or overweight adults, identified from the electronic patient database of a partnering healthcare organization. In particular, low-socioeconomic status individuals and men were targeted for oversampling.

2.3.1 Survey development

The main objectives during development were 1) to ensure accessibility to low-socioeconomic status individuals, and 2) to identify and describe the attributes of incentive designs which are plausible and relevant to policy and practice.

The initial draft of the survey instrument was reviewed during a listening session with program assistants and instructors associated with the Expanded Food and Nutrition Education Program (EFNEP), a nutrition education program aimed at low-socioeconomic status individuals. Subsequent drafts of the survey instrument were piloted and discussed during four focus groups at different locations in the recruitment area. The first focus group was recruited from the client group of an EFNEP program assistant, and the remaining focus groups were recruited from clinics of our research partner, Carilion Clinic, a healthcare organization in southwest Virginia.

In order to meet the first objective, survey instrument readability was improved in a number of ways. Descriptions of the weight loss intervention and the incentive design attributes were shortened, and an effort made to use simpler language. Pictograms and examples were added to the descriptions to enhance comprehension. Format, layout and flow of the questions were designed to improve readability and facilitate higher response rates.

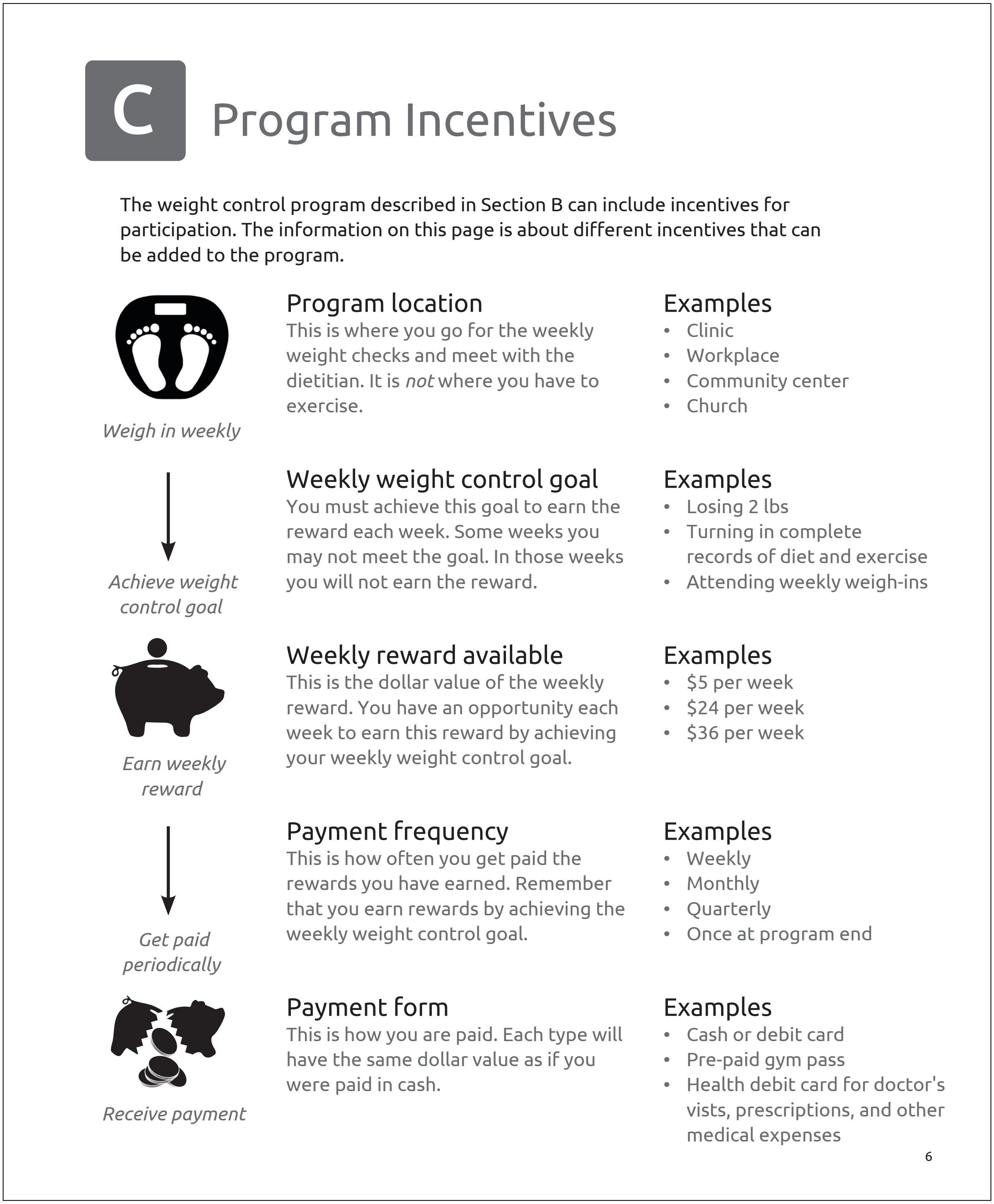
In order to meet the second objective, a literature review was conducted to form the initial set of attributes and attribute levels. Payment frequency, reward magnitude, payment form, and reward contingency were all attributes found to be varied in previous studies. From the focus groups, we learned that program location also had the potential to be an important attribute. A pilot study found that, for payment form, a grocery card was perceived in the same way as cash, so we replaced that the grocery card level with a health debit card that could be used for doctor's visits, prescriptions and other medical expenses.

Table 2.1 describes the final list of attributes and attribute levels used in our choice experiment, and Figure 2.1 displays the attributes and attribute levels as they were described in the survey instrument. Figure 2.2 displays an example of the choice task presented to respondents. Notice that respondents were allowed to opt out of the weight loss program (and perforce any accompanying incentive design).

The weight loss program with which all financial incentives were associated was a standard behavioural weight loss program (Pinto et al., 2007). The intervention was described as being six months long, and hence, the total reward magnitude attribute refers to the total over six months. The intervention involved a dietitian meeting, coaching calls, diet planning and tracking, exercise planning and tracking, and weekly weigh-ins (Figure 2.3).

Attribute	Attribute levels
Reward amount	\$0, \$48, \$96, \$216, \$384, \$576
Payment form	Cash Pre-paid gym pass (gym) *Health debit card for doctor's visits, prescriptions, and other medical expenses* (medical) Debit card (debit)
Payment frequency	*Once at end of program* (once) Quarterly Monthly Weekly
Program location	*Clinic* Workplace Community center Church
Reward condition	Losing 2 lbs (weight) *Attending weekly weight checks* (attendance) Turning in records of diet and exercise (compliance) Attending weekly weight checks and turning in records of diet and exercise (att.comp)

Table 2.1: Attributes and attribute levels used in constructing the financial incentive. Abbreviated terms used in the text are in parentheses. Baseline levels for categorical attributes are set in italics. Reward amount is defined as the total possible payment over the duration of the six-month intervention. Choice sets as presented to respondents included both the weekly reward amount and the total reward amount.



C

Program Incentives

The weight control program described in Section B can include incentives for participation. The information on this page is about different incentives that can be added to the program.

Weigh in weekly

Program location

This is where you go for the weekly weight checks and meet with the dietitian. It is *not* where you have to exercise.

Examples

- Clinic
- Workplace
- Community center
- Church

Achieve weight control goal

Weekly weight control goal

You must achieve this goal to earn the reward each week. Some weeks you may not meet the goal. In those weeks you will not earn the reward.

Examples

- Losing 2 lbs
- Turning in complete records of diet and exercise
- Attending weekly weigh-ins

Earn weekly reward

Weekly reward available

This is the dollar value of the weekly reward. You have an opportunity each week to earn this reward by achieving your weekly weight control goal.

Examples

- $5 per week
- $24 per week
- $36 per week

Get paid periodically

Payment frequency

This is how often you get paid the rewards you have earned. Remember that you earn rewards by achieving the weekly weight control goal.

Examples

- Weekly
- Monthly
- Quarterly
- Once at program end

Receive payment

Payment form

This is how you are paid. Each type will have the same dollar value as if you were paid in cash.

Examples

- Cash or debit card
- Pre-paid gym pass
- Health debit card for doctor's vists, prescriptions, and other medical expenses

6

Figure 2.1: Description of financial incentive attributes and attribute levels as presented in the survey instrument.

1 Please consider the following two weight control programs.

	Program A	Program B
Program location	Clinic	Church
Weekly weight control goal	Attending weekly weight checks	Turning in records of diet and exercise
Weekly reward available	$16	$4
Payment frequency	Once at end of program	Weekly
Payment form	Health debit card for doctor's visits, prescriptions, and other medical expenses	Debit card
Total reward available in program	$16 × 24 wks = $384	$4 × 24 wks = $96

Which weight control program would you choose?
Please check one box.

- ☐ 1 Program A
- ☐ 2 Program B
- ☐ 3 I would not choose either program.

114 10

Figure 2.2: Example of choice question

B Weight control program

This section is about a weight control program to start people on the path to losing weight. It is 6 months long, which is 24 weeks. People who followed this program closely lost about

1-2 lbs	→	24-48 lbs
per week	results in	over 24 weeks

Each picture below represents one part of the program.

Dietitian meeting
Have a one-time meeting with a dietitian to develop personal diet and exercise plans.

Coaching calls
Receive weekly calls to help you follow your diet and exercise plans.

Diet plan
Use a personal workbook with recipes and more to plan healthy meals.

Diet tracking
Record meals and snacks online or on paper. Find out how many calories you ate.

Exercise plan
Use a personal workbook to plan exercise that's just right for you.

Exercise tracking
Record physical activity online or on paper. Find out how many calories you used.

Program weigh-ins
Weekly weight checks at program location. Only program staff will see your weight.

4

Figure 2.3: Description of behavioural weight loss intervention associated with financial incentive design, as presented to respondents.

Choice sets (*i.e.*, the pair of incentive designs presented in the choice task) were constructed using a D-efficiency design, in which a large number of possible designs were drawn from the full factorial design and sequentially compared with respect to the D-efficiency criterion[1]. A total of 96 choice sets were constructed, and each survey instrument contained four choice sets, resulting in 24 instrument versions. These were assigned randomly and distributed as evenly as possible to the survey sample.

2.3.2 Survey implementation

Survey participants were recruited through our research partner, Carilion Clinic, using their electronic patient database. Based on eligibility criteria (adult obese or overweight) and desired demographic profile (*e.g.*, oversampling males and low socioeconomic status individuals, as inferred by insurance status), 7,554 patients were randomly drawn from the database. These patients were mailed introductory letters, signed by their primary care physician, describing the study and alerting them to the next step of the recruitment process, a telephone call seeking consent. Only 2,737 individuals consented to the survey during phone recruitment, primarily due to the high rate of invalid phone numbers in the electronic patient database (88% reachable by phone consented to the study). Those individuals who consented to the survey by telephone were sent the survey via mail, then a reminder postcard, and finally another copy of the survey if they had not yet responded. The number of completed surveys received was 1,297, yielding a 47% survey completion rate. Of the completed surveys, one survey was excluded because the respondent was younger than 18 years old. Hence, the final number of surveys analyzed was 1,296.

2.3.3 Supplementary data

As described in subsection 2.2.4, including demographic interactions and conducting analyses on predicted choice probabilities open the possibility for generalizing results beyond the study sample by using appropriate weights. We use the 2009-2010 National Health and Nutrition Examination Survey (NHANES) as a supplementary dataset to provide these weights. NHANES is a nationally representative survey which provides demographic and health data, including all of the relevant demographic variables identified in the methods section (race, gender, income, employment status and education). Each observation in the NHANES dataset is assigned a sample weight. To arrive at the sample weight for a specific combination of individual characteristics, we sum the weights for

[1]The D-efficiency criterion refers to the goal of maximizing the determinant of the Fisher information matrix of the parameter estimates.

all the observations which display that specific combination. When we predict a participation probability for a specific individual profile, we also assign the profile's sample weight to the participation probability. Then, we use the sample weights when aggregating across probabilities (*e.g.*, taking a weighted mean of the probabilities). This approach can be seen as creating a synthetic population based on the adult US population, and predicting market shares in that synthetic population (Hensher et al., 2005, p 440).

Note that the NHANES sample weights and generalization to sub-populations are only applied for analyses which use predicted participation probabilities. None of the discrete choice models are actually estimated using a weighted sample. It is only the participation probabilities predicted from the estimated primary conditional logit model which are weighted. The analysis of attribute level preferences uses predicted participation probabilities, while the whole attribute importance analysis relies on log-likelihood values of the estimated model, and therefore is limited to in-sample interpretation.

2.4 Results

We begin this section by describing the characteristics of the study's sample, which reveals that our oversampling strategy was successful in recruiting a more heterogeneous sample, better reflecting the sub-populations vulnerable to obesity and overweight. We use this heterogeneous sample to assess the presence of preference heterogeneity across individuals using a secondary random parameters logit model. Once we establish preference heterogeneity, we control for it using relevant demographic variables in the primary conditional logit model. We interpret the estimated primary conditional logit model by rank ordering whole attributes by their relative importance for each subgroup of interest. Attributes identified as important have a disproportionate impact on weight loss program participation decisions, and should therefore receive careful attention in future studies or implementations. We also interpret the estimated primary conditional logit model by assessing relative preferences for attribute levels within each attribute and for each individual profile. Attribute levels which are highly preferred will increase the probability of weight loss program participation the most, and should therefore be a part of future studies or implementations. Due to significant preference heterogeneity, the most preferred attribute levels are not necessarily the same for each individual, and so we propose and show the benefits of offering more than one level for each attribute.

2.4.1 Sample characteristics

Table 2.2 compares the socio-demographic characteristics of the respondent sample to those of Virginia. Compared to the rest of Virginia, our sample is relatively more African-American, male, low-income, unemployed, and overweight or obese. These differences show that our strategy of recruiting overweight or obese adults[2] and oversampling males and low socioeconomic status individuals was successful. Our sample is heterogeneous and reflects those most vulnerable to overweight and obesity or those who have been understudied in previous weight loss studies.

2.4.2 Preference heterogeneity

We first assess the presence of preference heterogeneity across individuals using the secondary random parameters logit model. Table 2.3 presents the coefficient estimates of the model, whose specification attribute level variables and an alternative-specific constant. The alternative-specific constant `ASC.SQ` represents the effect of the opt-out alternative, and can be interpreted as the remaining effect on utility the opt-out alternative has, after accounting for all the attribute levels. Another constant to note is the term `Mag.0`, which helps to differentiate the opt-out alternative from a design alternative with the \$0 reward amount. The only attribute difference between the two alternatives is program location, and so we include the constant `Mag.0` in order to further differentiate the two alternatives, beyond program location. Note that all discrete variables are effect-coded, and in particular, the two constants are binary variables which take value -1 when true and 1 when false. Therefore, a positive coefficient on `ASC.SQ` is interpreted as an overall willingness to participate in a weight loss program with a financial incentive. A positive coefficient on `Mag.0` is interpreted as increased willingness to participate when facing an incentive design with a non-zero reward amount. Also note that the only continuous variable, reward amount, is log-transformed[3].

The coefficient standard deviation estimates reveal significant preference heterogeneity across individuals. For each attribute, at least one level displays a standard deviation estimate significantly different from zero. When the standard deviation estimates are significant, the coefficient of variation is always greater than 1, and often much greater.

[2]Discrepancy between our eligibility criteria and respondent characteristics can be explained by the discrepancy between the source of data. Recruitment into the study was based on an electronic patient database, which included height and weight (but of unknown date and accuracy), whereas respondent characteristics were computed based on self-reported height and weight from the survey instrument.

[3]As suggested in Lancsar and Louviere (2008), the choice of transformation is based on graphing estimates of the coefficients on reward amount levels (in a conditional logit model) and observing a log-linear relationship between reward amount level and coefficient estimates. Details can be found in section 6.B.

	Sample	Virginia
Race		
White	49.0%	71.3%
Black	41.1%	19.8%
Native American	0.48%	0.1%
Asian	4.0%	5.8%
Gender		
Female	43.1%	50.9%
Income		
Household income $10,000 or less	17.6%	5.7%
Employment status		
Percent unemployed	42.3%	6.9%
Education		
High school graduate or higher, percent age 25+	87.0%	86.6%
Bachelor's degree or higher, percent age 25+	27.1%	34.4%
Weight status		
Underweight	0.1%	2.0%
Normal weight	7.2%	34.4%
Overweight	38.3%	36.2%
Obese	54.3%	27.4%

Table 2.2: Comparison of sample characteristics to Virginia demographic characteristics. Sources of data include Virginia QuickFacts (US Census Bureau), American Community Survey (2008-2012 5-year estimates), and the BRFSS prevalence and trends data (2012). Weight status is assigned based on computed BMI, using CDC definitions of weight classification.

	Means		Standard deviations	
ASC.SQ	0.076	(0.117)	1.455***	(0.078)
Reward amount				
log(Mag + 1)	0.377***	(0.043)	0.440***	(0.025)
Mag.0	0.038	(0.121)	−0.386**	(0.131)
Program location				
Workplace/Clinic	−0.227***	(0.048)	−0.116	(0.115)
Community center/Clinic	0.171***	(0.049)	0.314**	(0.106)
Church/Clinic	−0.193***	(0.050)	0.514***	(0.101)
Payment form				
gym/cash	−0.126*	(0.053)	0.549***	(0.113)
medical/cash	−0.181***	(0.051)	−0.321**	(0.117)
debit/cash	0.175***	(0.051)	0.171	(0.127)
Reward condition				
weight/attendance	0.118*	(0.054)	0.288*	(0.122)
compliance/attendance	−0.072	(0.052)	0.492***	(0.112)
att.comp/attendance	−0.143**	(0.049)	−0.201	(0.120)
Payment frequency				
weekly/once	0.131*	(0.057)	0.334**	(0.118)
monthly/once	−0.027	(0.049)	0.064	(0.127)
quarterly/once	0.042	(0.051)	0.202	(0.117)
Log-likelihood	−4362.491			
N	4994			

Table 2.3: Random parameters logit model estimates. Column 1 reports the estimated coefficient means while column 2 reports the estimated standard deviations. All coefficients were assumed to follow a normal distribution. Standard errors are reported in parentheses below the coefficient estimates. See the text for the meaning of the alternative-specific constants `ASC.SQ` and `Mag.0`.

These results indicate that we should attempt to capture some of this heterogeneity with observed characteristics. Earlier, we identified the relevant demographic variables as race, gender, income, employment status, education and weight status on the grounds of policy relevance. Hence, the primary conditional logit model we use for the remainder of the analysis includes attribute level variables and demographic variables as main effects as well as interaction terms between the attribute level variables and demographic variables.

Now that we have established preference heterogeneity, we estimate the primary conditional logit model, which captures that heterogeneity via interaction terms between attributes and demographic variables. Given the extensive number of coefficient estimates for this model and the difficulty of interpreting these estimates, we present them in section 6.A. Instead of considering the coefficient estimates directly, we interpret them in terms of attribute importance and, through predicted choice probabilities, attribute level preferences.

2.4.3 Attribute importance

We now consider the relative importance of whole attributes for each policy relevant demographic group. Figure 2.4 presents the results of the partial log-likelihood analysis. Relative importance in the figure is measured as the partial effect relative to the sum of all partial effects for a given sub-sample.

For the full sample, reward amount is predictably more important than any other attribute. However, program location is the second-most important, and for several subgroups (blacks, low-income women and males), program location is more important than reward amount. This result is consistent with our experiences during focus groups, where participants were particularly concerned with program location when discussing their choices (due to a number of reasons, such as convenience, travel time, and privacy).

The least important attributes were payment form, reward condition and payment frequency. However, their rankings among the subgroups vary greatly. The black subsample found reward condition more important and payment frequency least important, while low-income women found payment frequency more important and reward condition least important. One would expect that individuals concerned with payment frequency would also be concerned with reward amount, but for low-income women, the program location matters more. One possible reason for this ranking is that low-income women may be more likely to have constraints on their transportation options, and so program location may reflect a financial concern, rather than simply convenience.

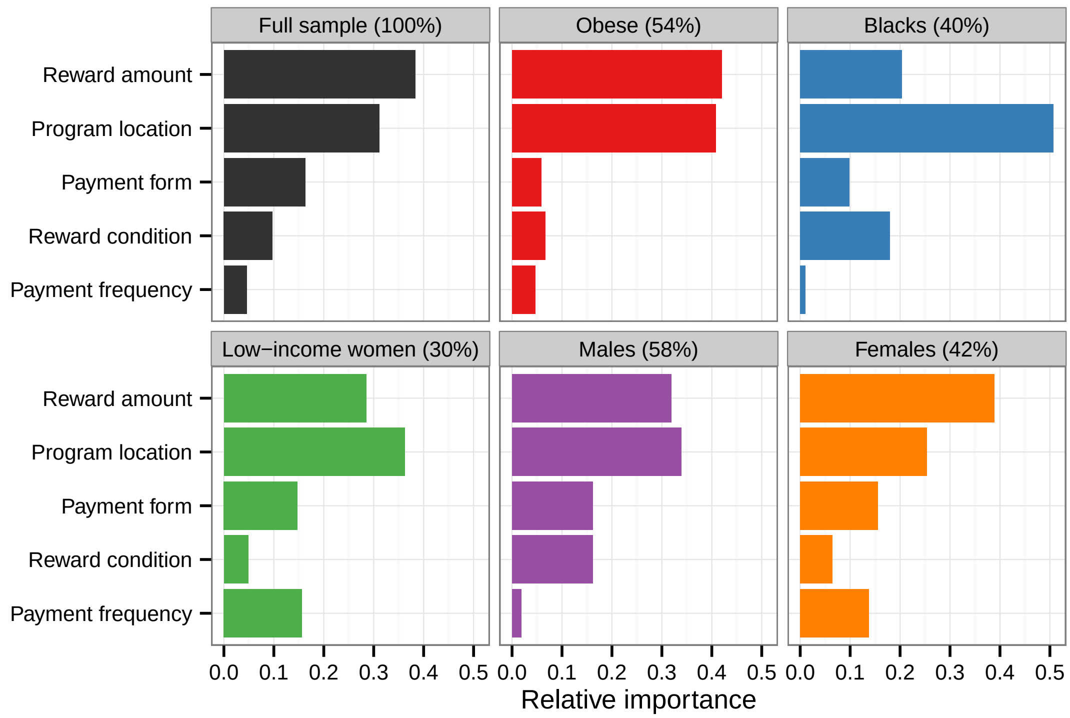


Figure 2.4: Relative importance of whole attributes as measured by partial log-likelihood analysis. Percentages next to each sub-sample name indicates the proportion of the full sample consisting of the sub-sample. 'Low-income women' refers to females in the sample who indicated an annual household income of $40,000 or less.

2.4.4 Reward amount

Since the most important attribute for most individuals is reward amount, we begin our discussion of attribute level preference with the reward amount attribute. Table 2.4 presents a subset of the estimates from the primary conditional logit model, including only the coefficients for terms involving reward amount. The first column contains estimates for the main effect of the log-transformed reward amount, as well as the interaction terms involving that variable. The sign of the estimate can be interpreted as the direction of influence and the magnitude as the responsiveness to reward amount. We observe that the overall grand mean effect is positive, but several of the interaction terms are negative, thereby decreasing that effect. If an individual possesses enough characteristics that are associated with negative interaction terms, then the overall effect of increasing reward amount on participation will be negative.

Since a negative preference for increasing reward amount is concerning from a policy

	`log(Mag+1)`		`Mag.0`	
Main effect	0.209***	(0.057)	0.063	(0.158)
Race interaction				
White	0.007	(0.057)	0.078	(0.159)
Black	−0.091	(0.059)	0.137	(0.163)
Hispanic	0.088	(0.096)	−0.094	(0.266)
Other	−0.004		−0.121	
Gender interaction				
Female	0.003	(0.034)	−0.091	(0.094)
Income interaction				
Very low	−0.082	(0.075)	0.278	(0.209)
Low	0.096*	(0.053)	−0.145	(0.149)
Medium	0.066	(0.054)	−0.352**	(0.148)
High	−0.080		0.220	
Employment status interaction				
Employed	−0.051	(0.034)	−0.017	(0.094)
Education interaction				
High school or less	−0.139***	(0.048)	0.282**	(0.133)
Some college or technical school	−0.024	(0.043)	0.109	(0.121)
College graduate or beyond	0.163		−0.390	
Weight status interaction				
Normal	−0.099	(0.079)	0.088	(0.219)
Overweight	−0.032	(0.052)	0.085	(0.144)
Obese	0.131		−0.173	

*** Significant at 1% level, ** significant at 5% level, * significant at 10% level.

Table 2.4: Coefficient estimates for main effect and interaction terms involving reward amount selected from the primary conditional logit model. The first column collects terms involving the log-transformed reward amount variable. The second column collects terms involving the binary variable `Mag.0` which is true when the design alternative has $0 reward amount (differentiating it from when the $0 reward amount is associated with the opt-out alternative). Standard errors are in parentheses. All categorical variables are effect-coded, hence the coefficient estimate for the baseline level of each categorical variable can be recovered, but the standard error cannot. Binary variables are also effect-coded, with -1 value when true and 1 value when false.

perspective, we examine how pervasive this preference may be. Table 2.5 summarizes sub-population shares which are predicted to have a negative response to increasing reward amount. Except for the African-American sub-population, most of the shares are small (about 10%), and in the case of the obese sub-population, particularly small (about 1%). These findings are consistent with the attribute importance results, in which reward amount was relatively less important to African Americans than to other subgroups.

National	Obese	Black	Low-income women	Males	Females
9.5%	0.9%	23.8%	9.8%	7.6%	11.3%

Table 2.5: Proportion of sub-population which is predicted to have a negative response to increasing reward amount.

Although decreasing willingness to participate with increasing reward amount seems counter-intuitive, the phenomenon matches focus group observations and previous literature. Some individuals in the focus groups reacted negatively to the presence of a financial reward, expressing the sentiment that individuals should lose weight for their health, not for financial gain. Financial incentives in behavioural health interventions have already raised a degree of ethical controversy (Halpern et al., 2009; Pearson and Lieber, 2009; Schmidt et al., 2010; Ashcroft, 2011; Madison et al., 2011; Reisinger et al., 2011; Schmidt, 2011; Schmidt et al., 2012; Klein, 2013; Lunze and Paasche-Orlow, 2013), and the acceptability of health incentives to the general public is relatively low (Promberger et al., 2011; Park et al., 2012). Health is not alone in this regard; the distaste for monetized transactions extends to a wide range of economic transactions (*e.g.*, organ donation). Roth (2007) describes three main reasons for such distaste, including objectification, coercion or exploitation, and a slippery slope. Investigating which reasons apply in the particular case of financial incentives in weight loss programs is beyond the scope of our study, but our finding of the negative impact of reward amount among some individuals suggests further study of distastefulness would be valuable, particularly among target populations and individuals who would actually be eligible for the incentives (rather than the general public).

The second column of Table 2.4 contains estimates for the main and interaction effects of `Mag.0`, which represents whether the design alternative has a $0 reward amount. A positive coefficient can be interpreted as an increased willingness to participate when offered a non-zero reward. Both positive and negative coefficients can be seen, indicating that some characteristics (such high school or less education) are associated with preferring a non-zero reward, but others (such as a medium level of income) are associated with preferring a zero reward. In some cases, the interaction effect of reward amount is the opposite direction from that of `Mag.0`. For example, high school or less education is associated with preferring a non-zero reward and a decreasing willingness to participate with increasing reward. This type of interaction effect, along with the possibility of overall negative effect of increasing reward amount, suggests that low reward amounts will be more effective at encouraging participation than high reward amounts. Low reward amounts will attract those who have a positive preference for increasing reward amount or non-zero rewards, while being more attractive to those who have a negative preference for increasing reward amount than a higher reward amount would be.

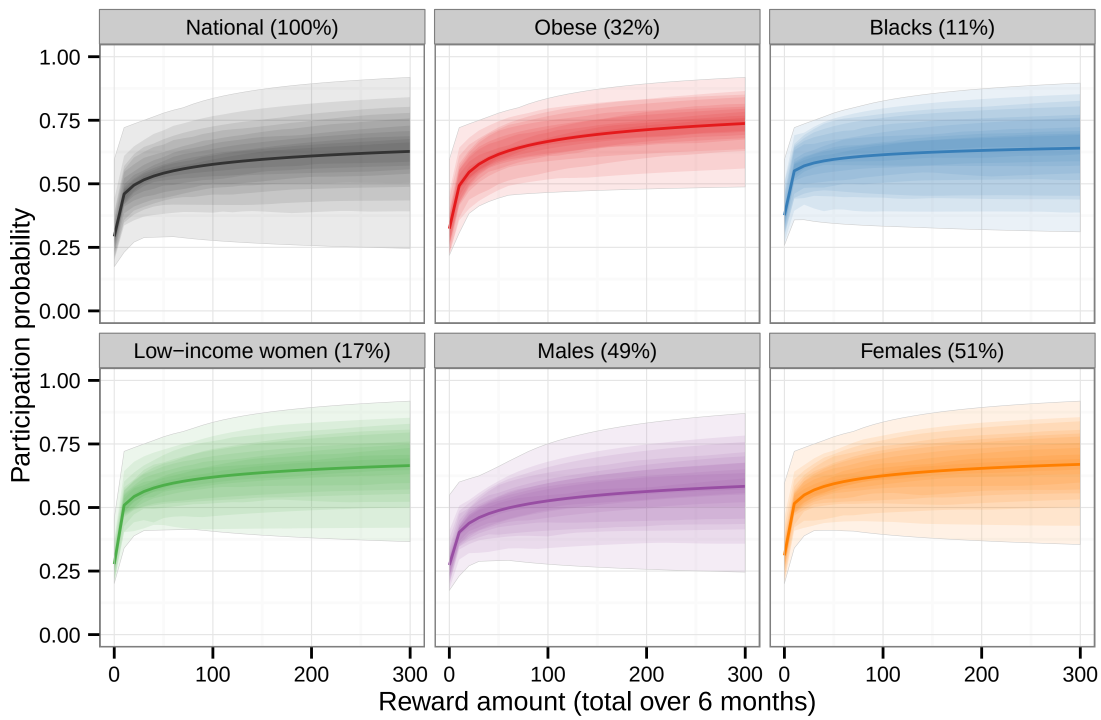


Figure 2.5: Response of participation probability to increasing reward amount for policy relevant demographic variables. Shaded regions indicate the distribution of probabilities, with darker colours indicating greater density. The dark line in the middle represents the weighted mean for that sub-population (using sample weights from NHANES). Note that only reward magnitude and the demographic variables contribute to utility. Percentages beside the sub-population names indicate the size of the sub-population relative to the full national population.

In order to gain a sense of what the range in responsiveness to the reward magnitude means in terms of participation probability, we predict participation probabilities at increasing reward amounts for each policy relevant demographic group (Figure 2.5). Participation probabilities are computed as the probability that an individual will not choose to opt out, and the predictions are weighted by NHANES sample weights in order to generalize the predicted probabilities to different sub-populations. For every sub-population, the mean response curve slopes upward, which is consistent with Table 2.5, showing that only small proportions of each sub-population have negative preferences for increasing reward amount. However, we can see those negative preferences in the bottom edge of the shaded region, which decreases with increasing reward amount for all sub-populations but the obese sub-population. This exception is consistent with Table 2.5, which indicates only an extremely small proportion of the obese sub-population is expected to have a negative preference for increasing reward amount.

The dispersion of the predicted probability distribution is greatest for the national population, which is no surprise, since it is the most heterogeneous. In contrast, the obese sub-population seems to be the most homogeneous, with the least amount of dispersion in the predicted probability distribution. The obese sub-population also displays the most responsiveness to increasing reward amount, in contrast to the African-American sub-population, which displays the least responsiveness. These observations are all consistent with Table 2.4 and Table 2.5, in which the African-American sub-population was predicted to have the highest proportion of individuals who have a negative preference towards increasing reward amount.

Finally, we make the important observation that the predicted probability increases fastest at the lowest reward amounts, a pattern that holds true across all sub-populations. This observation suggests again that programs should be designed with relatively low reward amounts, in order to attract the majority of individuals who would prefer an increasing reward amount, while minimizing the alienation of the minority who do not prefer an increasing reward amount.

2.4.5 Attribute level preferences

After reward amount, we consider the remaining attributes. In the series of figures below, we summarize the rankings of attribute levels within each attribute. As a reminder, the summaries are weighted by NHANES sample weights so that the results are representative of the sub-populations of interest, and the rankings are performed by predicted participation probability.

Under program location (Figure 2.6), most subgroups prefer the clinic and community centre (that is, those attribute levels have high rates of ranking first or second and low rates of ranking last), and the obese sub-population, in particular, prefer the clinic. In contrast, most subgroups do not prefer the church (that is, it has low rates of ranking first and high rates of ranking last). Workplace is an interesting mix between the other attribute levels, because it high rates of ranking first *and* last. This observation suggests

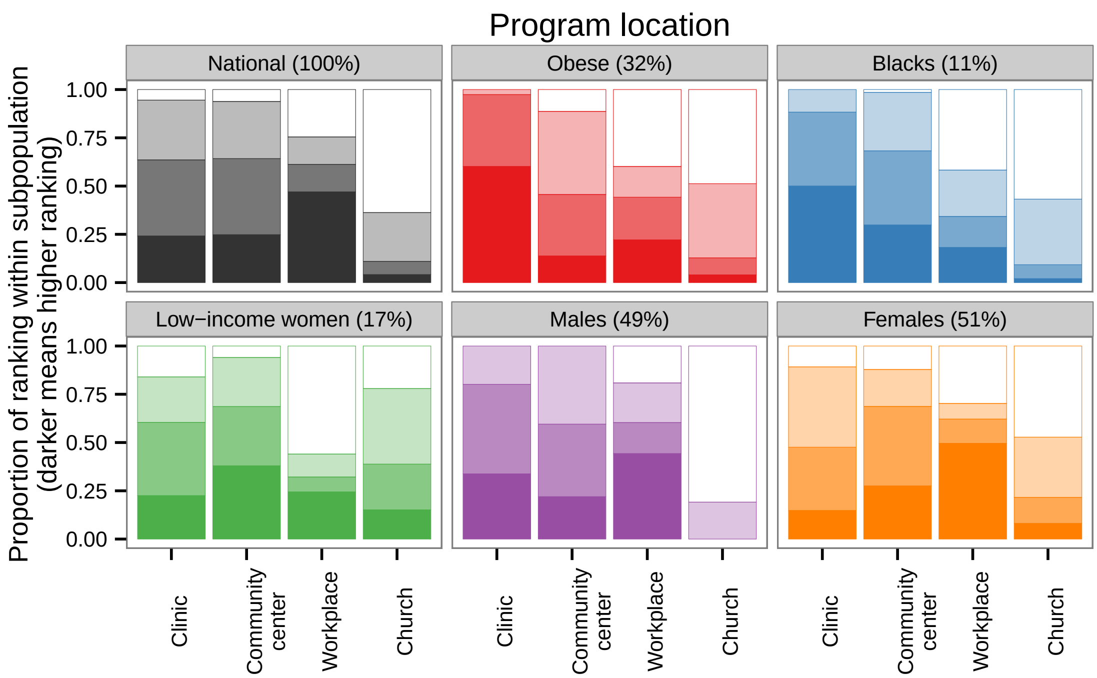


Figure 2.6: Rankings by attribute level for program location.

substantial heterogeneity in the attitude towards locating the program in the workplace, with some individuals greatly preferring it and others greatly preferring other locations. This heterogeneity is perhaps due to differing employment status; among low-income women, the unemployment rate is higher, and accordingly the rates of ranking workplace lower are much higher. On the other hand, among women overall, the rates of ranking workplace first are higher than among other sub-populations. The implication is that women prefer the workplace location when they're employed and but not when unemployed.

Under payment form (Figure 2.7), all subgroups clearly prefer the debit card, with nearly the entire male sub-population predicted to rank the debit card first. On the other hand, the gym pass has a higher rate of ranking first than the debit card among the African-American sub-population. This type of preference may be due to the same reasons we speculate drive the African-American negative preference for increasing reward amount: disapproval of using financial rewards for weight loss. Moreover, the gym pass most directly supports weight loss compared to the other payment forms, thus providing a possible explanation for why the gym pass is more preferred than the debit card for this sub-population. Finally, the medical card is clearly the least preferred payment form

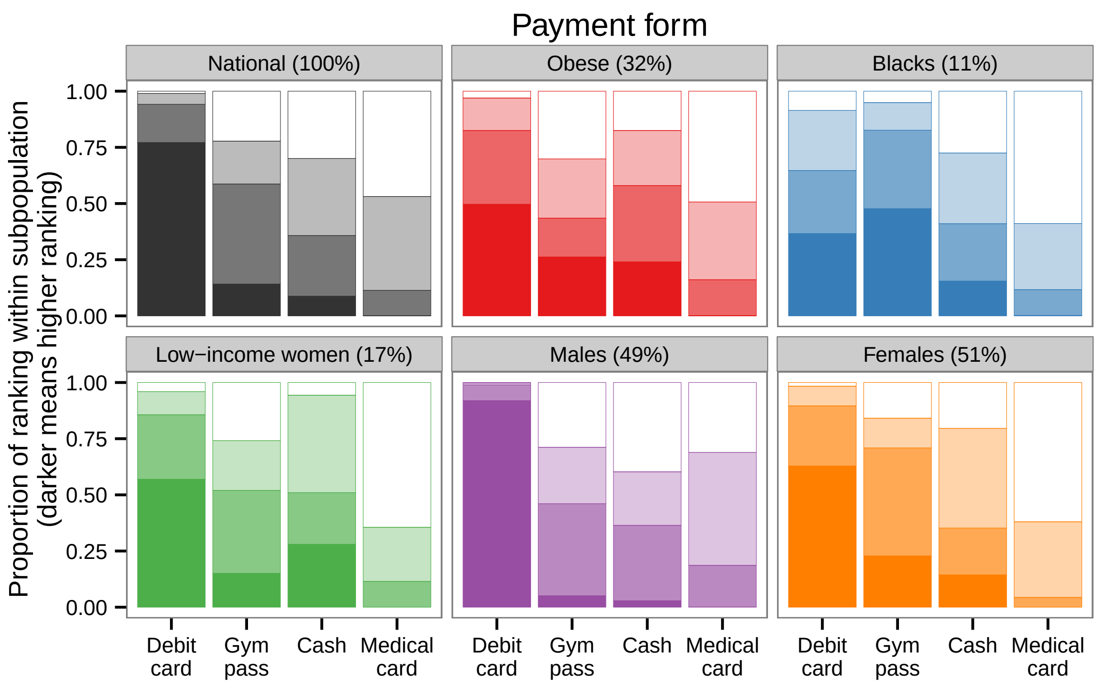


Figure 2.7: Rankings by attribute level for payment form.

among all subgroups.

Under reward condition (Figure 2.8), most subgroups prefer attendance and weight loss. These two attribute levels represent an interesting contrast, because they represent very different levels of effort and risk. On the one hand, weigh-in attendance requires low levels of effort and risk, whereas actual weight loss calls for high levels of effort and risk. However, weight loss is the ultimate objective of the weight loss program, and so it makes sense that many individuals would be attracted to this reward condition. Moreover, the much higher African-American preference for weight loss over weigh-in attendance is consistent with previous preferences emphasizing weight loss over financial reward. Surprisingly, between the two less favoured attribute levels, attending weigh-ins and keeping records is more favoured than solely keeping records. This result is surprising because combining attendance and record-keeping can only be more burdensome and risky than record-keeping alone.

Under payment frequency (Figure 2.9), most subgroups prefer weekly or quarterly payments. While quarterly payments had low rates of ranking last, weekly payments had high rates of ranking first. Why quarterly payments are relatively preferred is difficult

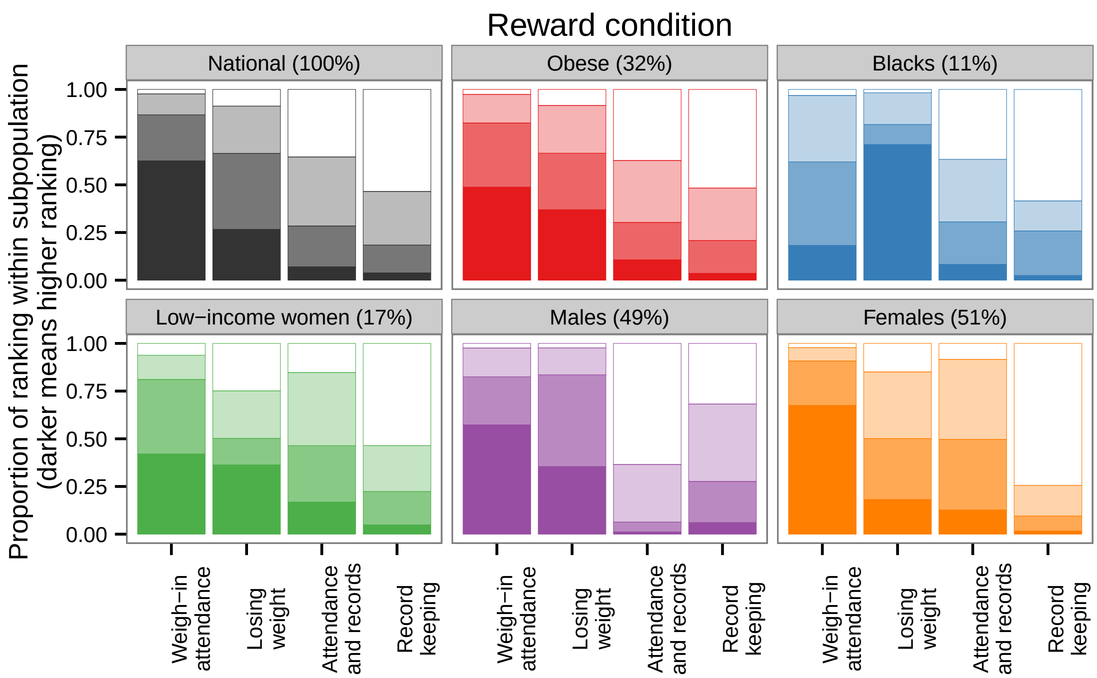


Figure 2.8: Rankings by attribute level for reward condition.

to say, especially since a one-time payment at the end of the program was clearly least preferred by all subgroups (although, the reader may not be surprised to see that the African-American sub-population had lower rates of ranking the one-time payment last compared to the other sub-populations, which is consistent with an expectation that a one-time payment may be more effective at incentivizing weight loss than more frequent payments).

We summarize the preferred attribute levels from the above discussion in Table 2.6, which displays the two most preferred attribute levels for each attribute.

2.4.6 Multiple incentive designs

The preference heterogeneity across individuals suggests that programs which offer more than one incentive design may achieve greater participation and representativeness. Different incentive designs appeal to different individuals, and offering more than one would allow interventions to appeal to more individuals. To investigate the return on offering multiple designs, we predict the participation probabilities under choice sets

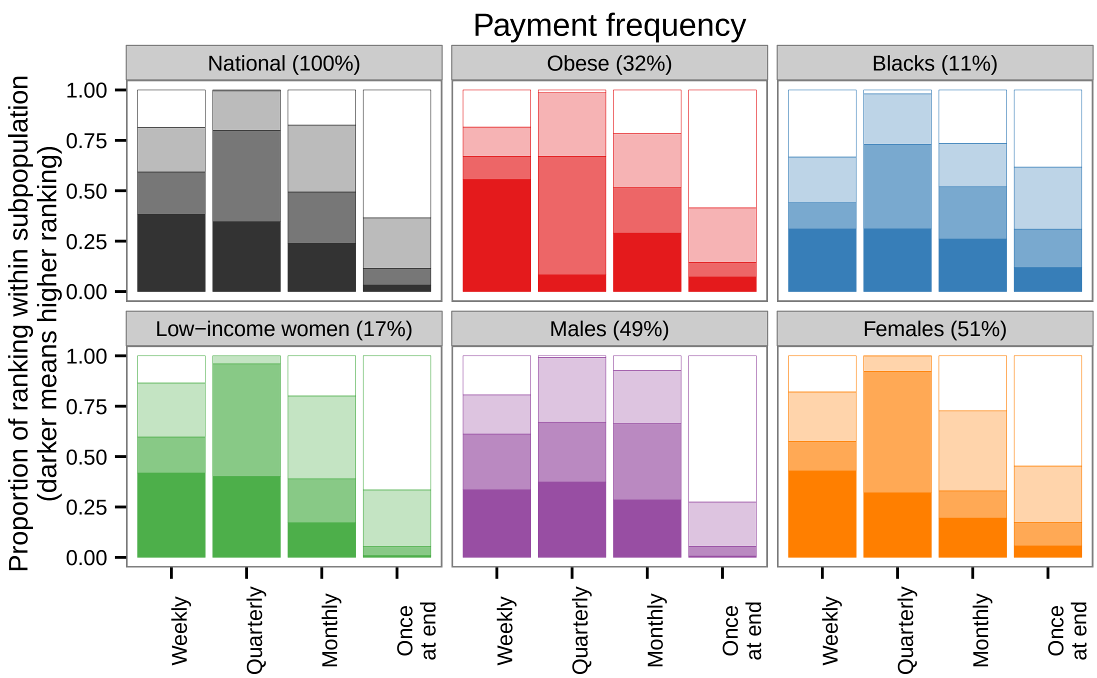


Figure 2.9: Rankings by attribute level for payment frequency.

including more than one design alternative. The design alternatives we use for this analysis come from the recommended attribute levels (Table 2.6). We construct five choice sets: 1) where only one design alternative is offered, corresponding to the first row of Table 2.6, 2) where two design alternatives are offered, allowing program location to take on the two values in the first column of Table 2.6, 3) where four design alternatives are offered, allowing both program location and payment form to take on two levels each, and so on until 5) where all discrete attributes are allowed to take on two levels each.

Program location	Payment form	Reward condition	Payment frequency
Clinic	Debit card	Attendance	Weekly
Community center	Gym pass	Weight loss	Quarterly

Table 2.6: Summary of recommended attribute levels.

Figure 2.10 presents the participation probability distributions for the national population in grey, and the weighted participation probability means by sub-population for each of the five choice sets. As more design alternatives are offered, the national participation probability distribution decreases in range, decreases in dispersion and increases in mean. Thus, giving individuals more choice has two benefits, of increasing overall participation probability and reducing heterogeneity in participation probability. At a fixed reward magnitude of $100 in this illustration, the obese sub-population starts off with the highest participation probability, whereas African Americans have the lowest predicted participation probability. But as the number of incentive designs increases, the African-American sub-population catch up, overtaking the male sub-population. However, by the maximum number of incentive designs in this illustration, the mean participation probability is nearly indistinguishable between sub-populations. Hence, increasing the number of designs offered would seem to be an effective method for countering preference heterogeneity.

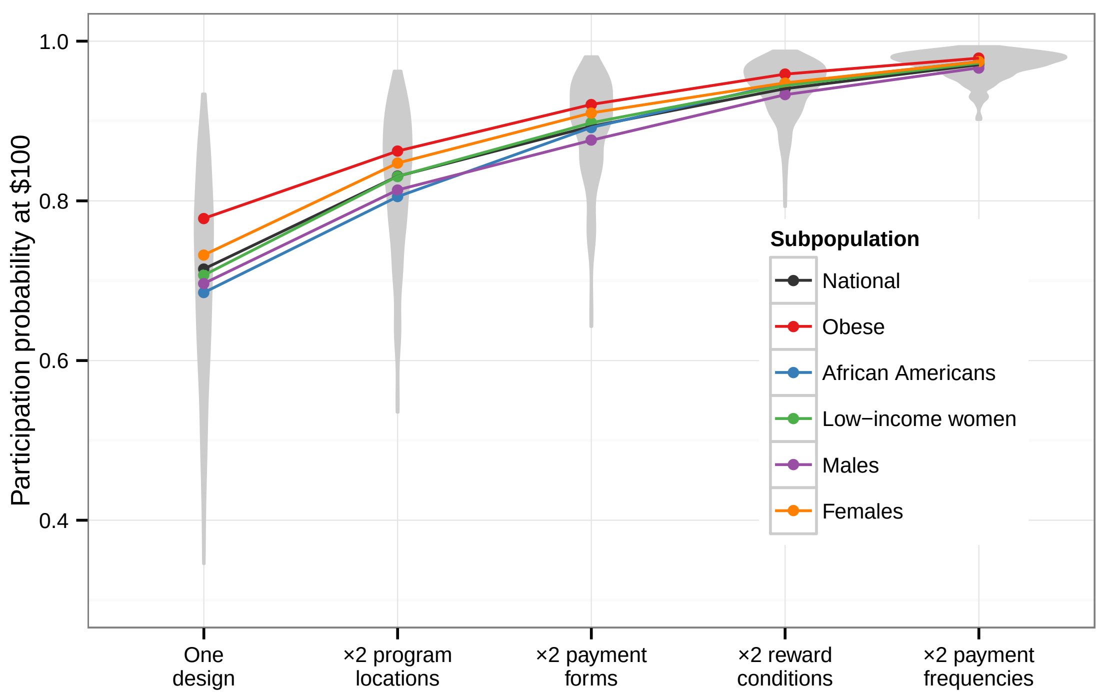


Figure 2.10: Participation probability distribution when multiple designs are presented. Each value on the x-axis represents a different choice set; the choice sets are progressively larger as more attributes are given two levels (see text for further detail). The grey shaded regions are violin plots (a kernel density plot which is rotated and mirrored) for the national population. The points and lines indicate sub-population-specific weighted mean.

2.5 Conclusion

Although financial incentives in weight loss programs have long been studied, their effect on participation and representativeness have rarely been examined. This study used a discrete choice experiment to design financial incentives which would maximize participation among target populations. We targeted populations vulnerable to obesity and overweight and which have been understudied in past studies of weight loss: African Americans, low-income women and males. The choice experiment presented financial incentives in terms of five attributes: reward amount, program location, payment form, reward condition and payment frequency.

We first established the presence of significant preference heterogeneity for each of the attributes, leading us to explicitly model the observed portions of this heterogeneity by including in the model specification interactions with policy relevant demographic variables: race, gender, income, employment status, education and weight status. A natural consequence of this preference heterogeneity is that offering more than one design alternative increases the probability of participation. We show that offering multiple design alternatives is an effective counter to preference heterogeneity, reducing the spread of the distribution of participation probabilities, and collapsing the participation probability means of the target groups until they are practically indistinguishable.

We found that reward amount and program location were the most important attributes for all populations. Even so, we found that having a negative preference for increasing reward amount was possible, and most probable in the African-American sub-population. In fact, we found many consistent signs that a disproportionate number of individuals in the African-American sub-population prefer non- or minimally-monetized incentives [4].

- Program location was a more important attribute than reward amount for the African-American sub-sample.
- The predicted proportion of negative preferences for increasing reward amount was highest, by far, for the African-American sub-population.
- African Americans were predicted to be the least responsive to increasing reward amount.
- Within the African-American sub-population, the gym pass payment form was more likely to be ranked ahead of the debit card payment form.

[4] Note that these observations and their implication do not arise from the race variable alone: since any analysis at the level of predicted choice probabilities is weighted to be representative of the African-American sub-population, these observations rely just as much on other demographic variables which differentiate the African-American sub-population from other populations.

- Within the African-American sub-population, the weight loss reward condition was more likely to be ranked ahead of the attendance condition.
- Compared to other sub-populations, the African-American sub-population was less likely to rank the one-time payment frequency last.

The obese subgroup, on the other hand, provides a stark contrast to the African-American subgroup. The obese subgroup appears to prize the monetary aspect of the financial incentive.

- Program location and reward amount were far more important than the other incentive attributes.
- The obese subgroup was highly responsive to the reward amount, with an extremely low rate of negative preference for increasing reward amount.
- The obese subgroup preferred the low effort and high reward attribute levels – weekly payments and the weigh-in attendance reward condition.

The relatively high rate of non-monetary motivation among the African-American subgroup suggests that offering low reward amounts (which do not alienate those with negative preferences towards increasing reward amounts) and multiple incentive designs (which gives individuals greater freedom to choose the incentive design which they prefer) are important strategies to maximize the participation among African Americans.

Maximizing participation among those who are already obese, on the other hand, is simpler because this subgroup is primarily concerned with the monetary aspect of the incentive. This subgroup is less responsive to the strategy of offering multiple designs because they are mostly focused on the reward amount – the higher, the better. However, offering multiple designs does no harm to the participation rates in this subgroup. The non-negative effect of this strategy makes it superior to the strategy of offering higher reward amounts, which would alienate those individuals who have a negative preference towards increasing reward amount.

Thus, choosing the reward amount requires a careful balance between those who respond positively and those who respond negatively. Ultimately, the specific reward amount depends upon policy objectives dictating which individuals are the target of the weight loss program. Choosing the levels of the other attributes is less fraught; although some subgroups have strong tendencies towards some levels (*e.g.*, clinic program location for the obese subgroup, debit card payment form for the male subgroup), all subgroups are fairly consistent in the top two levels for each discrete attribute (summarized in Table 2.6).

Future work could investigate the causes for the negative response to reward amount and possible approaches to ameliorating the negative response. Previous cases of repugnance have been successfully treated with in-kind rather than monetary transactions (Roth, 2007), and we already see evidence along this line in the African-American preference for the gym pass over the debit card payment form. In addition, if the payment form were not linked to a dollar amount, then framing effects could lead to different or more insightful results.

Future work could also investigate program location more closely, since it is the second-most important attribute and one which displays a great deal of heterogeneity. Program location itself could be decomposed into constituent attributes, which may provide a more coherent picture of location preferences. Understanding why some individuals prefer some locations over others can suggest locations not previously studied, new locations which could be built, or predict which existing location would be most suitable given a target population and geographic area.

Despite considering more incentive components than previous studies, there are still more ways in which financial incentives can vary. Evidence suggests that deposit contracts lower participation rates (*e.g.*, Jeffery et al., 1990), and future work could ascertain this result as well as check for individual heterogeneity. Escalating payments, in which later payments are larger than earlier payments, may increase effectiveness (*e.g.*, Stitzer and Petry, 2006), but their effect on participation is unknown. Adding uncertainty to the incentive design may also increase effectiveness (*e.g.* Cawley and Price, 2011), but again, the effect on participation is unknown. Previous studies suggest that group-based incentives may be more powerful than individual incentives (*e.g.*, Jeffery, 2012), but the effect on participation and the role of individual heterogeneity is unknown. Exploiting group dynamics may be an inexpensive way to increase reach and effectiveness.

Bibliography

Ashcroft, Richard E. Personal financial incentives in health promotion: where do they fit in an ethic of autonomy? *Health Expectations*, 14(2):191–200, 2011. ISSN 1369-7625. doi: 10.1111/j.1369-7625.2011.00664.x. URL http://dx.doi.org/10.1111/j.1369-7625.2011.00664.x.

Augurzky, Boris, Bauer, Thomas K., Reichert, Arndt R., Schmidt, Christoph M., and Tauchmann, Harald. Does money burn fat? Evidence from a randomized experiment. Ruhr Economic Papers 368, Ruhr-Universität Bochum (RUB), Department of Economics, September 2012.

Burns, Rachel J., Donovan, Angela S., Ackermann, Ronald T., Finch, Emily A., Rothman, Alexander J., and Jeffery, Robert WW. A theoretically grounded systematic review of material incentives for weight loss: Implications for interventions. *Annals of Behavioral Medicine*, 44:375–388, 2012. ISSN 0883-6612. doi: 10.1007/s12160-012-9403-4. URL http://dx.doi.org/10.1007/s12160-012-9403-4.

Butsch, W. Scott, Ard, Jamy D., Allison, David B., Patki, Amit, Henson, C. Suzanne, Rueger, Miriam M., Hubbert, Katherine A., Glandon, Gerry L., and Heimburger, Douglas C. Effects of a reimbursement incentive on enrollment in a weight control program. *Obesity*, 15(11):2733–2738, 2007. ISSN 1930-739X. doi: 10.1038/oby.2007.325. URL http://dx.doi.org/10.1038/oby.2007.325.

Cameron, R, MacDonald, MA, Schlegel, RP, Young, CI, Fisher, SE, Killen, JD, Rogers, T, Horlick, L, and Shepel, LF. Toward the development of self-help health behaviour change programs: weight loss by correspondence. *Can J Public Health*, 81(4):275–279, Jul-Aug 1990.

Cawley, John and Price, Joshua A. *Outcomes in a Program that Offers Financial Rewards for Weight Loss*, pages 91–126. University of Chicago Press, April 2011. URL http://www.nber.org/chapters/c11816.

Cawley, John and Price, Joshua A. A case study of a workplace wellness program that offers financial incentives for weight loss. *Journal of Health Economics*, 32(5): 794–803, 2013. ISSN 0167-6296. doi: 10.1016/j.jhealeco.2013.04.005. URL http://www.sciencedirect.com/science/article/pii/S0167629613000544.

Crane, Melissa M., Tate, Deborah F., Finkelstein, Eric A., and Linnan, Laura A. Motivation for participating in a weight loss program and financial incentives: an analysis from a randomized trial. *J Obes*, 2012:290589, 2012. doi: 10.1155/2012/290589. URL http://dx.doi.org/10.1155/2012/290589.

Finkelstein, Eric A, Linnan, Laura A, Tate, Deborah F, and Birken, Ben E. A pilot study testing the effect of different levels of financial incentives on weight loss among overweight employees. *J Occup Environ Med*, 49(9):981–989, Sep 2007. doi: 10.1097/JOM.0b013e31813c6dcb. URL http://dx.doi.org/10.1097/JOM.0b013e31813c6dcb.

Flegal, KM, Carroll, MD, Kit, BK, and Ogden, CL. Prevalence of obesity and trends in the distribution of body mass index among us adults, 1999–2010. *JAMA*, 307(5):491–497, 2012. doi: 10.1001/jama.2012.39. URL http://dx.doi.org/10.1001/jama.2012.39.

Forster, JL, Jeffery, RW, Sullivan, S, and Snell, MK. A work-site weight control program using financial incentives collected through payroll deduction. *J Occup Med*, 27(11): 804–8, 1985. URL http://www.ncbi.nlm.nih.gov/pubmed/4067685.

Glasgow, RE, Vogt, TM, and Boles, SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *Am J Public Health*, 89(9):1322–1327, September 1999. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1508772/.

Halpern, Scott D., Madison, Kristin M., and Volpp, Kevin G. Patients as mercenaries?: The ethics of using financial incentives in the war on unhealthy behaviors. *Circulation: Cardiovascular Quality and Outcomes*, 2(5):514–516, 2009. doi: 10.1161/CIRCOUTCOMES.109.871855. URL http://circoutcomes.ahajournals.org/content/2/5/514.short.

Hensher, D.A., Rose, J.M., and Greene, W.H. *Applied Choice Analysis: A Primer*. Applied Choice Analysis: A Primer. Cambridge University Press, 2005. ISBN 9780521605779.

Heshka, S, Anderson, JW, Atkinson, RL, Greenway, FL, Hill, JO, Phinney, SD, Kolotkin, RL, Miller-Kovach, K, and Pi-Sunyer, FX. Weight loss with self-help compared with a structured commercial program: A randomized trial. *JAMA*, 289(14):1792–1798, 2003. doi: 10.1001/jama.289.14.1792. URL http://dx.doi.org/10.1001/jama.289.14.1792.

Holmes, Thomas P. and Adamowicz, Wiktor L. *A Primer on Nonmarket Valuation*, chapter Attribute-Based Methods, pages 171–219. Kluwer Academic Publishers, Dordrecht, 2003.

Hubbert, K A, Bussey, B F, Allison, D B, Beasley, T M, Henson, C S, and Heimburger, D C. Effects of outcome-driven insurance reimbursement on short-term weight control. *International Journal of Obesity*, 27(11):1423–1429, 2003. ISSN 0307-0565. URL http://dx.doi.org/10.1038/sj.ijo.0802403.

Jeffery, R W and French, S A. Preventing weight gain in adults: design, methods and one year results from the pound of prevention study. *International Journal of Obesity & Related Metabolic Disorders*, 21(6):457, 1997. ISSN 03070565. URL http://www.nature.com/ijo/journal/v21/n6/pdf/0800431a.pdf.

Jeffery, R W and French, S A. Preventing weight gain in adults: the pound of prevention study. *Am J Public Health*, 89(5):747–751, 1999. doi: 10.2105/AJPH.89.5.747. URL http://ajph.aphapublications.org/cgi/content/abstract/89/5/747.

Jeffery, R W, Forster, J L, French, S A, Kelder, S H, Lando, H A, McGovern, P G, Jacobs, D R, and Baxter, J E. The healthy worker project: a work-site intervention for weight control and smoking cessation. *Am J Public Health*, 83(3):395–401, Mar 1993a. ISSN 0090-0036. URL http://dx.doi.org/10.2105/AJPH.83.3.395.

Jeffery, Robert W. Financial incentives and weight control. *Preventive Medicine*, 55, Supplement:S61–S67, 2012. ISSN 0091-7435. doi: 10.1016/j.ypmed.2011.12.024. URL http://www.sciencedirect.com/science/article/pii/S0091743511005263.

Jeffery, Robert W., Thompson, Paul D., and Wing, Rena R. Effects on weight reduction of strong monetary contracts for calorie restriction or weight loss. *Behaviour Research and Therapy*, 16(5):363–369, 1978. ISSN 0005-7967. doi: 10.1016/0005-7967(78)90005-0. URL http://www.sciencedirect.com/science/article/pii/0005796778900050.

Jeffery, Robert W., Gerber, Wendy M., Rosenthal, Barbara S., and Lindquist, Ruth A. Monetary contracts in weight control: Effectiveness of group and individual contracts of varying size. *Journal of Consulting and Clinical Psychology*, 51(2):242–248, 1983. ISSN 1939-2117(Electronic);0022-006X(Print). doi: 10.1037/0022-006X.51.2.242.

Jeffery, Robert W., Bjornson-Benson, Wendy M., Rosenthal, Barbara S., Kurth, Candace L., and Dunn, Mary M. Effectiveness of monetary contracts with two repayment schedules on weight reduction in men and women from self-referred and population samples. *Behavior Therapy*, 15(3):273–279, 1984. ISSN 0005-7894. doi: 10.1016/S0005-7894(84)80029-5. URL http://www.sciencedirect.com/science/article/pii/S0005789484800295.

Jeffery, Robert W., Forster, Jean L., and Snell, Mary Kaye. Promoting weight control at the worksite: A pilot program of self-motivation using payroll-based incentives. *Preventive Medicine*, 14(2):187–194, 1985. ISSN 0091-7435. doi: 10.1016/0091-7435(85)90034-9. URL http://www.sciencedirect.com/science/article/pii/0091743585900349.

Jeffery, Robert W., Hellerstedt, Wendy L., and Schmid, Thomas L. Correspondence programs for smoking cessation and weight control: A comparison of two strategies in the Minnesota Heart Health Program. *Health Psychology*, 9(5):585–598, 1990. ISSN 1930-7810(Electronic);0278-6133(Print). URL http://psycnet.apa.org/journals/hea/9/5/585/.

Jeffery, Robert W, Wing, Rena R, Thorson, Carolyn, Burton, Lisa R, Raether, Cheryl, Harvey, Jean, and Mullen, Monica. Strengthening behavioral interventions for weight loss: a randomized trial of food provision and monetary incentives. *Journal of consulting and clinical psychology*, 61(6):1038, 1993b. URL http://psycnet.apa.org/journals/ccp/61/6/1038.html.

Jeffery, Robert W., Wing, Rena R., Thorson, Carolyn, and Burton, Lisa R. Use of personal trainers and financial incentives to increase exercise in a behavioral weight-loss program. *Journal of Consulting and Clinical Psychology*, 66(5):777–783, 1998. ISSN 1939-2117(Electronic);0022-006X(Print). URL http://psycnet.apa.org/doi/10.1037/0022-006X.66.5.777.

John, Leslie K., Loewenstein, George, Troxel, Andrea B., Norton, Laurie, Fassbender, Jennifer E., and Volpp, Kevin G. Financial incentives for extended weight loss: A randomized, controlled trial. *Journal of General Internal Medicine*, 26(6):621–626, 2011.

ISSN 0884-8734. doi: 10.1007/s11606-010-1628-y. URL http://dx.doi.org/10.1007/s11606-010-1628-y.

Klein, Eran P. Patient health incentives: Ethical challenges and frameworks. *International Journal of Behavioral Medicine*, pages 1–10, November 2013. ISSN 1070-5503. doi: 10.1007/s12529-013-9373-3. URL http://dx.doi.org/10.1007/s12529-013-9373-3.

Kramer, F. Matthew, Jeffery, Robert W., Snell, Mary Kaye, and Forster, Jean L. Maintenance of successful weight loss over 1 year: Effects of financial contracts for weight maintenance or participation in skills training. *Behavior Therapy*, 17(3):295–301, 1986. ISSN 0005-7894. doi: 10.1016/S0005-7894(86)80058-2. URL http://www.sciencedirect.com/science/article/pii/S0005789486800582.

Kullgren, Jeffrey T., Troxel, Andrea B., Loewenstein, George, Asch, David A., Norton, Laurie A., Wesby, Lisa, Yuanyuan, Tao, Jingsan, Zhu, and Volpp, Kevin G. Individual-versus group-based financial incentives for weight loss. *Annals of Internal Medicine*, 158 (7):505–514, 2013. ISSN 00034819. URL http://annals.org/article.aspx?articleid=1671710.

Lahiri, Supriya and Faghri, Pouran D. Cost-effectiveness of a workplace-based incentivized weight loss program. *Journal of Occupational and Environmental Medicine*, 54(3): –, 2012. ISSN 1076-2752. URL http://journals.lww.com/joem/Fulltext/2012/03000/Cost_Effectiveness_of_a_Workplace_Based.20.aspx.

Lancaster, Kelvin J. A new approach to consumer theory. *Journal of Political Economy*, 74 (2):pp. 132–157, 1966. ISSN 00223808. URL http://www.jstor.org/stable/1828835.

Lancsar, Emily and Louviere, Jordan. Conducting discrete choice experiments to inform healthcare decision making. *PharmacoEconomics*, 26(8):661 – 677, 2008. ISSN 11707690. URL http://link.springer.com/article/10.2165/00019053-200826080-00004.

Lancsar, Emily, Louviere, Jordan, and Flynn, Terry. Several methods to investigate relative attribute impact in stated preference experiments. *Social Science & Medicine*, 64(8):1738 – 1753, 2007. ISSN 0277-9536. doi: http://dx.doi.org/10.1016/j.socscimed.2006.12.007. URL http://www.sciencedirect.com/science/article/pii/S0277953606006368.

Lunze, Karsten and Paasche-Orlow, Michael K. Financial incentives for healthy behavior: Ethical safeguards for behavioral economics. *American Journal of Preventive Medicine*, 44 (6):659 – 665, 2013. ISSN 0749-3797. doi: http://dx.doi.org/10.1016/j.amepre.2013.01.035. URL http://www.sciencedirect.com/science/article/pii/S0749379713001864.

Madison, Kristin M., Volpp, Kevin G., and Halpern, Scott D. The law, policy, and ethics of employers' use of financial incentives to improve health. *The Journal of Law, Medicine*

& Ethics, 39(3):450–468, 2011. ISSN 1748-720X. doi: 10.1111/j.1748-720X.2011.00614.x. URL http://dx.doi.org/10.1111/j.1748-720X.2011.00614.x.

Mahoney, Michael J. Self-reward and self-monitoring techniques for weight control. *Behavior Therapy*, 5(1):48–57, 1974. ISSN 0005-7894. doi: 10.1016/S0005-7894(74)80085-7. URL http://www.sciencedirect.com/science/article/pii/S0005789474800857.

Mann, Ronald A. The behavior-therapeutic use of contingency contracting to control an adult behavior problem: weight control. *Journal of Applied Behavior Analysis*, 5(2): 99–109, 1972. ISSN 1938-3703. doi: 10.1901/jaba.1972.5-99. URL http://dx.doi.org/10.1901/jaba.1972.5-99.

Mavis, Brian E. and Stöffelmayr, Bertram E. Multidimensional evaluation of monetary incentive strategies for weight control. *The Psychological Record*, 44(2):229–252, 1994.

Moller, Arlen C., McFadden, H. Gene, Hedeker, Donald, and Spring, Bonnie. Financial motivation undermines maintenance in an intensive diet and activity intervention. *Journal of Obesity*, 2012:–, 2012. doi: 10.1155/2012/740519. URL http://www.hindawi.com/journals/jobes/2012/740519.

Ogden, CL, Lamb, MM, Carroll, MD, and Flegal, KM. Obesity and socioeconomic status in adults: United States 2005–2008. NCHS data brief 50, National Center of Health Statistics, Hyattsville, MD, 2010. URL http://www.cdc.gov/nchs/data/databriefs/db50.pdf.

Pagoto, Sherry L., Schneider, Kristin L., Oleski, Jessica L., Luciani, Juliana M., Bodenlos, Jamie S., and Whited, Matt C. Male inclusion in randomized controlled trials of lifestyle weight loss interventions. *Obesity*, 20(6):1234–1239, 2012. ISSN 1930-739X. doi: 10.1038/oby.2011.140. URL http://dx.doi.org/10.1038/oby.2011.140.

Park, James D., Metlay, Jessica, Asch, Jeremy M., and Asch, David A. The New York Times readers' opinions about paying people to take their medicine. *Health Education & Behavior*, 39(6):725–731, 2012. doi: 10.1177/1090198111428645. URL http://heb.sagepub.com/content/39/6/725.abstract.

Paul-Ebhohimhen, V. and Avenell, A. Systematic review of the use of financial incentives in treatments for obesity and overweight. *Obesity Reviews*, 9(4):355–367, 2008. ISSN 1467-789X. doi: 10.1111/j.1467-789X.2007.00409.x. URL http://dx.doi.org/10.1111/j.1467-789X.2007.00409.x.

Pearson, Steven D. and Lieber, Sarah R. Financial penalties for the unhealthy? Ethical guidelines for holding employees responsible for their health. *Health Affairs*, 28(3): 845–852, 2009. doi: 10.1377/hlthaff.28.3.845. URL http://content.healthaffairs.org/content/28/3/845.abstract.

Petry, Nancy M., Barry, Danielle, Pescatello, Linda, and White, William B. A low-cost reinforcement procedure improves short-term weight loss outcomes. *The American Journal of Medicine*, 124(11):1082–1085, 2011. ISSN 0002-9343. doi: 10.1016/j.amjmed.2011.04.016. URL http://www.sciencedirect.com/science/article/pii/S0002934311003251.

Pinto, Angela Marinilli, Gokee-LaRose, Jessica, and Wing, Rena R. Behavioral approaches to weight control: a review of current research. *Women's Health*, 3(3):341–53, May 2007. ISSN 1745-5057. doi: 10.2217/17455057.3.3.341.

Promberger, Marianne, Brown, Rebecca C H, Ashcroft, Richard E, and Marteau, Theresa M. Acceptability of financial incentives to improve health outcomes in UK and US samples. *Journal of Medical Ethics*, 37(11):682–687, 2011. doi: 10.1136/jme.2010.039347. URL http://jme.bmj.com/content/37/11/682.abstract.

Reisinger, Heather Schacht, Brackett, Rachel Horner, Buzza, Colin D., Páez, Monica B. Williams, Gourley, Ryan, Vander Weg, Mark W., Christensen, Alan J., and Kaboli, Peter J. "All the money in the world..." patient perspectives regarding the influence of financial incentives. *Health Services Research*, 46(6pt1):1986–2004, 2011. ISSN 1475-6773. doi: 10.1111/j.1475-6773.2011.01287.x. URL http://dx.doi.org/10.1111/j.1475-6773.2011.01287.x.

Relton, Clare, Strong, Mark, and Li, Jessica. The 'pounds for pounds' weight loss financial incentive scheme: an evaluation of a pilot in NHS Eastern and Coastal Kent. *Journal of Public Health*, 33(4):536–542, 2011. doi: 10.1093/pubmed/fdr030. URL http://jpubhealth.oxfordjournals.org/content/33/4/536.abstract.

Reuss-Borst, Monika, Paloyo, Alfredo R., Reichert, Arndt R., and Tauchmann, Harald. Heterogeneous causal effects of financial incentives on weight loss: Results of a large prospective randomized trial. Ruhr Economic Papers 383, Rheinisch-Westfälisches Institut für Wirtschaftsforschung, Essen, 2012. URL http://hdl.handle.net/10419/67136.

Roth, Alvin E. Repugnance as a constraint on markets. *Journal of Economic Perspectives*, 21(3):37–58, 2007. doi: 10.1257/jep.21.3.37. URL http://www.aeaweb.org/articles.php?doi=10.1257/jep.21.3.37.

Ryan, Mandy and Gerard, Karen. Using discrete choice experiments to value health care programmes: current practice and future research reflections. *Applied Health Economics and Health Policy*, 2(1):55–64, 2003. URL http://eprints.soton.ac.uk/19109/.

Saccone, Anthony J. and Israel, Allen C. Effects of experimenter versus significant other-controlled reinforcement and choice of target behavior on weight loss. *Behavior Ther-*

apy, 9(2):271–278, 1978. ISSN 0005-7894. doi: 10.1016/S0005-7894(78)80112-9. URL http://www.sciencedirect.com/science/article/pii/S0005789478801129.

Schmidt, Harald. Wellness incentives, equity, and the 5 groups problem. *Am J Public Health*, 102(1):49–54, November 2011. ISSN 0090-0036. doi: 10.2105/AJPH.2011.300348. URL http://dx.doi.org/10.2105/AJPH.2011.300348.

Schmidt, Harald, Voigt, Kristin, and Wikler, Daniel. Carrots, sticks, and health care reform – problems with wellness incentives. *New England Journal of Medicine*, 362(2): e3, 2010. doi: 10.1056/NEJMp0911552. URL http://www.nejm.org/doi/full/10.1056/NEJMp0911552. PMID: 20042747.

Schmidt, Harald, Asch, David A., and Halpern, Scott D. Fairness and wellness incentives: What is the relevance of the process-outcome distinction? *Preventive Medicine*, 55, Supplement(0):S118 – S123, 2012. ISSN 0091-7435. doi: 10.1016/j.ypmed.2012.03.005. URL http://www.sciencedirect.com/science/article/pii/S0091743512000874.

Sherwood, Nancy E, Morton, Noelle, Jeffery, Robert W, French, Simone A, Neumark-Sztainer, Dianne, and Falkner, Nicole H. Consumer preferences in format and type of community-based weight control programs. *American Journal of Health Promotion*, 13 (1):12–18, 1998. doi: 10.4278/0890-1171-13.1.12.

Stitzer, Maxine and Petry, Nancy. Contingency management for treatment of substance abuse. *Annual Review of Clinical Psychology*, 2(1):411–434, 2006. doi: 10.1146/annurev.clinpsy.2.022305.095219. URL http://www.annualreviews.org/doi/abs/10.1146/annurev.clinpsy.2.022305.095219. PMID: 17716077.

Volpp, Kevin G., John, Leslie K., Troxel, Andrea B., Norton, Laurie, Fassbender, Jennifer, and Loewenstein, George. Financial incentive-based approaches for weight loss. *JAMA: The Journal of the American Medical Association*, 300(22):2631–2637, 2008. doi: 10.1001/jama.2008.804. URL http://jama.ama-assn.org/content/300/22/2631.abstract.

Wang, Youfa, Beydoun, May A., Liang, Lan, Caballero, Benjamin, and Kumanyika, Shiriki K. Will all Americans become overweight or obese? Estimating the progression and cost of the US obesity epidemic. *Obesity*, 16(10):2323–2330, 2008. ISSN 1930-739X. doi: 10.1038/oby.2008.351. URL http://dx.doi.org/10.1038/oby.2008.351.

WHO Consultation on Obesity. Obesity: preventing and managing the global epidemic: report of a WHO consultation. WHO technical report series 894, World Health Organization, Geneva, 2000. URL http://www.who.int/iris/handle/10665/42330.

Wing, Rena R., Epstein, Leonard H., Marcus, Marsha, and Shapira, Barbara. Strong monetary contingencies for weight loss during treatment and maintenance. *Behavior*

Therapy, 12(5):702–710, 1981. ISSN 0005-7894. doi: 10.1016/S0005-7894(81)80141-4. URL http://www.sciencedirect.com/science/article/pii/S0005789481801414.

Chapter 3

Improving the reliability of self-reported attribute non-attendance behaviour through the use of polytomous attendance scales

Abstract

The literature in discrete choice modelling is increasingly recognizing the existence of attribute non-attendance, in which respondents ignore some attributes when answering an attribute-based question. This behaviour may present a serious problem for modelling and inference, because it violates fundamental assumptions of the random utility model on which choice models are based. In this study, we elicit attribute non-attendance using a six-point polytomous attendance scale, rather than restricting them to a dichotomous ignored/considered response, as in previous studies. Stated non-attendance has been found to be unreliable in previous studies, but polytomous attendance scales have the potential to address the sources of unreliability. Using data from a choice experiment in health economics, this study assesses the performance and consistency between empirical observations and theoretical expectations of a polytomous attendance scale. We find that the lowest point on the attendance scale is the part of the scale which corresponds best to attribute non-attendance, and that attendance scales longer than two or three points do not provide much additional information. Furthermore, the polytomous attendance scale had limited success in producing theoretically consistent results, suggesting that potential for polytomous attendance scales to produce more reliable attendance statements was not realized in this study.

3.1 Introduction

For decades, stated preference methods have been used to value non-market goods and services. Today, one of the most popular approaches to stated preference is attribute-based methods, which conceptualize non-market goods or services as bundles of attributes, and estimate the value of each attribute. Although the theoretical and econometric foundations for attribute-based methods were developed decades ago, attribute-based methods are still undergoing improvement. One such area of active investigation is attribute non-attendance.

Attribute non-attendance refers to the phenomenon wherein respondents ignore some attributes when responding to an attribute-based question. The phenomenon has been empirically observed in self-reported attribute attendance (*e.g.* Hensher et al., 2005; Alemu et al., 2013) across transportation economics (*e.g.* Hensher, 2008), environmental economics (*e.g.* Campbell et al., 2008) and health economics (*e.g.* Hole, 2011). Attribute non-attendance may occur for many reasons: it may be an attempt to lessen choice task complexity, particularly by using heuristic rather than rational decision-making processes (Hensher, 2009; Payne et al., 1988), or it may occur because of attribute unimportance, unrealistic attribute levels and protest against trading off one attribute against others (Alemu et al., 2013; Carlsson et al., 2010).

Attribute non-attendance is a concern because it weakens the foundation on which attribute-based methods stand. In economics, the theoretical foundation of attribute-based methods is Lancaster's theory of characteristics of goods (Lancaster, 1966), which supposes that preference rankings are over the characteristics of the goods rather than the goods themselves. Hence, utility is specified as a function of characteristics rather than goods. McFadden (1974) developed the random utility model by uniting Lancaster's theory with random utility maximization, specifying utility in two parts, 1) systematic utility, which is deterministic and defined as a function of attributes, and 2) an error component, which endows random utility with its stochastic structure. The random utility model implies compensatory behaviour: increasing one attribute should be able to compensate for a decrease in another attribute. If individuals ignore some attributes, however, then they no longer display compensatory behaviour. Returning to basic consumer theory, such individuals have neither complete nor continuous preferences, and hence their preferences cannot be represented by a utility function. Consequently, maintaining theoretical consistency implies removing these individuals from the analysis.

However, a reliable method for identifying individuals with attribute non-attendance behaviour is still elusive. Early studies used self-reports of attendance behaviour, but later studies found that inconsistencies between self-reported attendance behaviour and estimated preferences. In theory, attribute non-attendance should correspond to zero marginal utility, but in many cases, non-zero marginal utility was estimated for attributes reportedly ignored. Furthermore, when studies attempted to infer attendance behaviour from responses to attribute-based questions, they found that inferred attendance frequently differed from stated attendance. As a result, self-reports of attendance behaviour are viewed as unreliable.

Polytomous attendance scales may be one way of eliciting more reliable self-reports of attendance behaviour. Nearly all previous stated attendance studies have utilized dichotomous attendance scales, where respondents are forced to choose between having 'attended' or 'ignored' an attribute. However, the dichotomous scale may be subject to unreliability due to misreporting, conflation between low and zero preferences, and choice task specific attendance. Polytomous attendance scales have the potential to address these sources of unreliability by lessening social desirability bias which leads to misreporting, distinguishing between low and zero preferences at finer scales, and specifying frequency of non-attendance across attribute-based questions.

In this study, we assess whether a six-point polytomous attendance scale fulfills the promise of improving reliability of self-reported attendance behaviour. We identify portions of the attendance scale which may correspond to zero marginal utility, investigate whether the original attendance scale is more informative than artificially shortened scales, and assess consistency between stated attendance levels and theoretical expectations and inferred attendance behaviour. The novel elicitation format is implemented in a choice experiment of financial incentive designs in the context of a behavioural weight loss program.

3.2 Literature review

Although attribute non-attendance and other types of non-compensatory decision-making has long been recognized in fields outside of economics (*e.g.* Payne et al., 1988), acknowledgement of this issue has come late to the choice modelling literature. Among the first to identify and address attribute non-attendance in choice modelling was Hensher et al. (2005), which directly elicited self reports of attribute attendance behaviour. Self-reported attribute non-attendance was assumed to correspond to zero marginal utility, and therefore coefficients on reportedly ignored attributes were constrained to zero in the model specification.

Early studies followed Hensher et al. (2005) in taking attribute attendance statements at face value, and assuming that stated non-attendance corresponded to zero marginal utility (Hensher, 2006; Hensher et al., 2007; Campbell et al., 2008; Puckett and Hensher, 2008). Most studies taking this approach found that imposing the assumption of zero marginal utility on ignored attributes resulted in better model fit than the conventional assumption of full attendance (Campbell et al., 2008; Campbell and Lorimer, 2009; Kaye-Blake et al., 2009; Meyerhoff and Liebe, 2009; Balcombe et al., 2011; Hess, 2012; Kehlbacher et al., 2013; Kosenius, 2013; Kragt, 2013), although not all studies found this improvement (Hensher et al., 2007; Carlsson et al., 2010; Rose et al., 2012; Alemu et al., 2013). These results, while not unequivocal, demonstrated the presence of attribute non-attendance and the importance of taking it into account.

Campbell and Lorimer (2009) was perhaps the first study to question this blind reliance on attribute non-attendance statements. This study checked for, rather than imposing, consistency between stated attendance behaviour and stated preference behaviour by allowing separate coefficients to be estimated for ignored and attended attributes. They found that, contrary to theory, some coefficients on ignored attributes were significantly different from zero. Following the same approach of separate coefficients, several additional studies also found that coefficient estimates on reportedly ignored attributes were significant (Campbell and Lorimer, 2009; Hess, 2012; Alemu et al., 2013; Hess and Hensher, 2013; Scarpa et al., 2013). Still, even when significant, coefficients on ignored attributes were sometimes found to be lower than attended attributes (Hess, 2012; Hess and Hensher, 2013), which led the authors to suggest that perhaps non-attendance is being conflated with low attendance. On the other hand, Carlsson et al. (2010) found no significant differences between coefficients on ignored and attended attributes. These studies suggest that attendance statements may not be reliable, because respondents who report attribute non-attendance may not have actually ignored those attributes.

Consequently, much of the literature is devoted to inferring attendance from choice behaviour alone, without the use of attendance statements. Various methods have been proposed to infer attendance, but review of those methods is beyond the scope of this study. To give the reader a flavour of the techniques employed to infer attendance, however, we briefly describe an approach proposed by Hess and Hensher (2010), and subsequently used by Mariel et al. (2012), Mariel et al. (2013) and Scarpa et al. (2013). In this approach, once a random parameters logit model has been estimated, attendance is inferred based on the respondent-specific coefficient distribution conditional on the respondent's observed choices. The idea is that attender and non-attenders will differ in the properties of their conditional distributions. In their empirical application, Hess and Hensher (2010) employed the following rule: if a conditional distribution has a coefficient of variation of 2 or greater, then the attribute is considered non-attended. The authors admit that this threshold is somewhat arbitrary, and in their simulation study, Mariel et al. (2013) found that the optimal threshold depends on the extent of non-attendance in the population.

Although model fit normally improves when accounting for attribute non-attendance, whether stated or inferred, there is very little agreement between stated and inferred attendance otherwise. Some studies have compared stated and inferred attendance at the respondent level: Hess and Hensher (2010) found that 10% to 50% of the respondents display a discrepancy between their stated attendance state and their inferred attendance, and Collins et al. (2013) found that 75% to 90% of their sample was 'aligned', although agreement between inferred non-attendance and stated non-attendance in particular ranged from 2% to 75% depending on the attribute and model. Other studies have compared inferred and stated attendance at the attendance profile level: Campbell et al. (2011) found the difference between rates of stated and inferred non-attendance range from 10 to over 50 percentage points. Overall, stated non-attendance rates tended to be higher than inferred non-attendance rates. Depending on the model and the attribute, Hole et al. (2013) found that differences ranged from less than 1 to more than 60 percentage points. For most models, inferred non-attendance rates were higher than stated non-attendance rates. Scarpa et al. (2013) found that most differences were less than 15 percentage points, but the differences were much larger for the full attendance profile. Inferred full attendance was much lower than stated full attendance, and consequently inferred non-attendance tended to be higher than stated non-attendance. Starting with all possible classes in an equality constrained latent class model, Kragt (2013) performed a specification search which only retained some of those classes. Of the retained classes, only about half corresponded to stated attendance profiles. Of those inferred classes which aligned with stated attendance profiles, there was a large difference in profile frequencies, from less than 5 percentage points to almost 40. Similar to Scarpa et al. (2013), Kragt (2013) found that inferred rates of full attendance was much lower than stated rates of full attendance. In addition to comparing attendance rates, Hess and Hensher (2010) also estimated models with separate coefficients for each attendance state, and found that the models using inferred attendance outperformed those using stated attendance. Mariel et al. (2012) performed the same type of analysis (although they estimated models with ignored coefficients constrained to zero rather than separate coefficients) and found the same result. Scarpa et al. (2011) used a multivariate probit model to explain the posterior probability of attendance for each attribute while including stated attendance as a covariate. They found that stated attendance was never significant when the statements came from a question asking which attributes respondents 'ignored', but were sometimes significant when the statements instead came from a question asking which attributes 'guided' the respondents' choices. Although there are large variations within studies, substantial differences between inferred attendance stated attendance appear to be common in most studies which compare them.

The discrepancies between stated attendance behaviour, stated preference behaviour and inferred attendance behaviour have cast doubt on the reliability of attendance statements. There are several possible reasons for this unreliability, including 1) misreporting, 2) conflation between low and zero preferences, and 3) choice task attendance.

In the conclusion to their seminal paper, Hensher et al. (2005) argue that the fundamental weakness of self-reports is that people do not always say what they think (*i.e.*misreport), either because they don't want to reveal it or because they don't actually know it (*i.e.*, the 'conscious-unconscious divergence'). Consequently, measuring the processes of decision-making successfully necessitates measuring unconscious attitudes and preferences. In other words, rather than relying on stated or inferred attendance, a third source of information could be *implied* attendance. In the decade since Hensher et al. (2005), only two studies have attempted to bridge the gap between conscious and unconscious behaviour. Kaye-Blake et al. (2009) conducted a choice experiment via computer-aided survey, in which respondents had to actively choose to reveal attribute level information by clicking on blank boxes. This design obviated the need to ask for attendance statements or to infer attendance – instead, attribute attendance was implied by which attributes respondents chose to reveal. The study found that the best performing model was not the 'accessed information model', which constrained to zero coefficients on attributes which respondents did not reveal, but rather the 'average information model', which set unrevealed attributes to their 'average' levels. This finding could have different interpretations, but one is that respondents may assume a baseline level when they 'ignore' attributes. The level range compared to this baseline may be small enough that their cognitive effort is best optimized by ignoring that attribute, even though they still have some non-zero marginal utility for the attribute and its baseline level. Balcombe et al. (2014b) used visual attention, as measured by eye-tracking, to imply attribute attendance. They found little correlation between visual attendance and stated attendance, but both were informative in model specifications, indicating a complementary nature between them. In short, studies which use attendance information implied by unstated behaviours reveal that misreporting may be common.

A special case of misreporting is conflation between low and zero preferences or between low and no attendance. If an individual has a low, yet non-zero preference for an attribute, then he may pay little attention to it and report non-attendance to it. Futhermore, inferred attendance methods cannot distinguish between low and zero preferences and this conflation may be responsible for some of the discrepancy between stated and inferred attendance. Attribute non-attendance is a special case of preference heterogeneity, and therefore the two are confounded. Among others, Hensher et al. (2013) has pointed out that inferred attendance models may include low preferences among zero preferences. Thus, conflation between low and zero preferences may affect stated attendance directly, or indirectly by reducing the consistency between inferred attendance and stated attendance.

Attribute attendance may vary across choice tasks (called choice task attendance in the literature), but most stated attendance studies implicitly assume constant attendance or non-attendance (called serial attendance in the literature) by virtue of eliciting attendance only once. A few studies have elicited attendance after every choice task, allowing respondents to indicate that they attended to some attributes in some choice tasks but not others. Meyerhoff and Liebe (2009) and Scarpa et al. (2010) found that choice task attendance was more common than serial attendance, and that accounting for choice task attendance resulted in better model fit than accounting for serial attendance. In fact, Scarpa et al. (2010) found that accounting for serial attendance actually worsened model fit compared to not accounting for attendance at all. Colombo and Glenk (2014) found that not only did models using choice task attendance have better fit than those using serial attendance, serial attendance was best reconstructed by assuming a 75% threshold for choice task attendance. That is, serial non-attendance was assigned when choice task non-attendance was reported for 75% or more of the choice tasks, rather than 100% as would be expected. These studies give the impression that serial attendance is the exception rather than the rule, thus opening up the vast majority of stated attendance studies to question.

Polytomous attendance scales, in contrast to the dichotomous attendance scales used in almost all stated attendance studies, have the potential to address all three sources of unreliability. Polytomous attendance scales, which allow respondents to mark different degrees of attendance, may ameliorate the effect of social desirability bias in misreporting, by allowing respondents to express some degree of non-attendance without actually having to indicate that they totally ignored an attribute. Polytomous attendance scales may also help to separate low and zero preferences, by providing attendance information at finer scales. Finally, polytomous attendance scales may capture some level of choice task attendance, since attendance can be expressed in terms of frequency (*e.g.*, none of the time, some of the time, all of the time). Eliciting attendance after each choice task would provide more information, in particular allowing attendance state to be compared with attribute levels, but multiple attendance questions is more burdensome for the respondent than a single debriefing question. In summary, polytomous attendance scales have the potential to provide more reliable attendance statements at relatively low implementation costs and low levels of respondent burden.

Polytomous attendance scales are a relatively understudied topic in the attribute non-attendance literature. Scarpa et al. (2013) and Weller et al. (2014) collected attendance based on a five-point scale, but recoded attendance responses to a binary scale for use in estimations. Neither study provided detailed justification for using the latter or a formal comparison between the two types of coding. Scarpa et al. (2010) briefly employed a polytomous scale during their pilot, but in their final study chose to use a dichotomous scale instead, and, similarly, detailed justification for the decision was not provided. Balcombe et al. (2014a) used attribute importance ranking data as coefficient mean shifters and scalers. Although ranking data is not dichotomous, it does not quite provide the same type of information as a polytomous attendance scale (since whether the least important attributes were ignored is not clear); furthermore, the usefulness of ranking data was not compared with dichotomous data. The only study to date which has used a polytomous attendance scale and compared its results with a dichotomous scale is Colombo et al. (2013). In their study, the authors used a trichotomous scale (*e.g.*, 'always considered', 'sometimes considered' and 'never considered'), which they collapsed to a dichotomous scale for some models but not for others. They found that models in which partial attendance was grouped with full attendance outperformed those where partial attendance was grouped with non-attendance, and coefficients separately estimated for ignored attributes were all insignificant when non-attendance was not collapsed into other groups. Furthermore, when separate coefficients were estimated for partial attendance and full attendance, both had the same pattern of significant coefficients, but the partial attendance coefficients were smaller (in absolute terms) than the full attendance coefficients. In short, every result regarding non-attendance and partial attendance was consistent with theoretical expectations, providing evidence that polytomous attendance scales can indeed produce more reliable attendance statements.

In this study, we contribute to the attribute non-attendance literature by providing a detailed comparison of polytomous attendance scale and a dichotomous scale recoded from it, and by using a longer, six-point polytomous attendance scale. We investigate which points on the polytomous attendance scale correspond to zero marginal utility, whether the original six-point scale can be recoded to shorter scales without losing significant information, and whether the original six-point scale and the recoded shorter scales result in coefficient estimates consistent with theoretical expectations.

3.3 Methods

In this section we describe the three categories of choice models we use to assess the polytomous attendance scale. The first category is the baseline, in which full attendance is assumed as in most conventional choice experiments. The second category constrains ignored attributes to zero marginal utility, thereby helping to determine which points on the polytomous attendance scale correspond to attribute non-attendance. The third category permits separate coefficients for attendance levels defined by shortened scales, allowing us to compare the information content and reliability of polytomous attendance scale with shorter scales. Using this third category model with separate coefficients for each attendance level, we assess the consistency of stated attendance with theoretical expectations. We also assess the consistency between stated and inferred attendance.

Choice experiment data are typically analyzed with choice models, which can be derived from the random utility model (Holmes and Adamowicz, 2003). The random utility model is based on the theory of characteristics of goods (Lancaster, 1966), in which preferences are defined over the attributes of the goods rather than the goods themselves. Specifically, the random utility an individual i receives from choosing an alternative j is defined as

$$U_{ij} = V_{ij} + \epsilon_{ij} \tag{3.1}$$

that is, the sum of a deterministic component, the systematic utility V_{ij}, and a stochastic component, the error term ϵ_{ij}.

Choice is modelled by comparing the random utility of different alternatives for the same individual. That is, an individual i is assumed to choose alternative j if the random utility of that alternative, U_{ij}, exceeds the random utility of all other alternatives in the choice set, $U_{im}, m \neq j$. Since utility is random rather than deterministic, choice is stochastic and discussed in terms of probabilities. The probability that an individual i chooses alternative j from a choice set is $p_{ij} = \Pr(U_{ij} > U_{im}) \forall m \neq j$. The distribution of the error term determines the type of choice model. In this work, we specify the error term to be iid Gumbel, resulting in the multinomial logit (MNL) model, which has the following choice probability expression

$$p_{ij} = \frac{\exp(V_{ij})}{\sum_m \exp(V_{im})}. \tag{3.2}$$

In order to avoid confounding the baseline attribute levels with the status quo levels, we employ effect coding (Holmes and Adamowicz, 2003; Hoyos, 2010; Bech and Gyrd-Hansen, 2005). Using effect coding, we can recover the effects on the baseline levels by taking the negative sum of the coefficients on the non-baseline levels. Recovering effects when considering interacted effect-coded variables is not quite as straightforward, but is still a matter of taking sums of coefficients (for more detail, see section 6.C). Thus, we can approximate standard errors for the recovered effects by using the delta method. In all presentations of MNL model coefficient estimates, we present recovered effects and approximated standard errors along with directly estimated coefficients and their standard errors.

We utilize a linear specification for the systematic utility,

$$V_{ij} = X_{ij}\beta \tag{3.3}$$

where X represents the attributes and their levels. We work with different specifications of β and X_{ij} to investigate different interpretations of the polytomous attendance scale. These specifications fall into three broad categories, 1) a baseline model, 2) non-attendance as zero marginal utility and 3) separate coefficients for each attendance level.

3.3.1 Baseline model

First, we will estimate a baseline model to which other model specifications can be compared. The natural baseline model is the full attendance model, which does not use any attribute non-attendance information. Hence, it makes the conventional assumption of full attendance to all attributes. In this model, each attribute is entered linearly into the systematic utility.

3.3.2 Non-attendance as zero marginal utility

We assess which points on the attendance scale are most consistent with the theoretical expectation of zero marginal utility for ignored attributes by estimating a set of models with coefficients on 'ignored' attributes constrained to zero. The models in the set will differ by which parts of the six-point polytomous attendance scale are considered to indicate that an attribute has been 'ignored'. For example, does the lowest point on the attendance scale correspond to 'ignore', as suggested in Colombo et al. (2013), or is it the lowest two points which correspond to 'ignore' as suggested in Scarpa et al. (2013)? Based on the wording of the elicitation, we expect only the lowest point to correspond to 'ignore', but we can empirically assess this expectation by estimating models under different specifications.

To make up this set of model specifications, we group adjacent points on the full attendance scale into the 'ignored' level in every possible way. For example, we could set 1) attendance level 1 to 'ignored', 2) attendance level 2 to 'ignored', 3) attendance levels 1 and 2 to 'ignored', 4) attendance levels 2 and 3 to 'ignored', and so on until every possibility is covered. Note that for this set of models, the vector of specified coefficients are not changing from model to model. Rather what is changing is the logic driving which attributes at the individual level are constrained to zero marginal utility.

The best-fitting model in this set will suggest to us the point(s) on the attendance scale with the most face validity as corresponding to 'ignoring' attributes. To select the best-fitting models, we will employ log-likelihood values and non-nested testing. We use log-likelihood values because each model has the same number of estimated parameters and the same null likelihood. Consequently, comparing information criteria (*e.g.*, AIC, BIC) and goodness-of-fit measures (*e.g.*, McFadden's pseudo R^2) is tantamount to comparing vanilla log-likelihood values. However, relying on log-likelihood values alone only tells us which models are best, but does not tell us if they are significantly better than the other models. In order to make probabilistic statements about which models are better than others, we will employ the Vuong test for non-nested hypotheses (Vuong, 1989), described in greater detail in section 6.D. In order to minimize the problems inherent in multiple comparisons, we will limit our tests only to a few of the top-ranked models, as ordered by log-likelihood value.

3.3.3 Attendance scale length

As mentioned previously, most attribute non-attendance studies use dichotomous, or, in a rare exception, trichotomous attendance scales. However, there may be benefits to using a longer scale, such as the six-point polytomous scale used in this study. To assess whether a longer scale is beneficial, we recode scales so that they are shorter than six points and compare them to the full six-point scale. These comparisons will help us determine how many points are necessary in an attendance scale.

The scales are recoded by assigning adjacent points to the same attendance level. For example, we can construct a recoded dichotomous scale by assigning points 1–3 to one attendance level and points 4–6 to another attendance level. In fact, there are $\binom{5}{1}$ ways of recoding to a dichotomous scale from the original six-point scale, allowing only adjacent points to be assigned to the same attendance level. Further, we can recode longer scales, from two levels up to five levels, in the same manner. In total there are $\sum_{l=2}^{5} \binom{5}{l-1} = 30$ configurations.

For each of the thirty recoded scales, we estimate a model with separate coefficients for each attendance level. Note that for this set of models, the specified vector of coefficients changes with the number of attendance levels, since a separate coefficient is estimated for each attendance level. However, for models with the same number of attendance levels, the specified vector of coefficients remains the same between models; what is different is the logic driving the construction of attendance levels, *i.e.*, which points on the six-point attendance scale correspond to a given attendance level.

This specification attempts to take into account attribute attendance information without imposing theoretical expectations on the data. The goal is to assess the relative amount of information available from each scale and, in particular, to determine whether longer scales are more informative than shorter scales. If longer scales are more informative, then there is justification for using them rather than the more common dichotomous scale. A significant difference between a model estimated under a shorter scale and the original six-point scale would suggest that the six-point scale has additional information not captured by the shorter scale, and so the longer scale may be more desirable. On the other hand, the lack of a significant difference implies that a shorter scale would suffice, and that burdening respondents with a longer scale may be unnecessary.

Because every configuration of a shorter scale is nested within the original six-point scale, we can use the likelihood ratio test for nested hypotheses. As in the previous section, we minimize multiple comparisons by considering top-ranked models according to model selection criteria. Unlike the previous section, the models may differ in the number of estimated parameters, and so we use measures which adjust for the degrees of freedom (*e.g.*, AIC, BIC and $\bar{\rho}^2$), rather than vanilla log-likelihood values. Due to the policy relevance of the predicted choice probabilities themselves (as representations of potential market uptake), we also use prediction accuracy as a model selection criterion. In this study, prediction accuracy is measured using K-fold cross validation with a multinomial deviance loss function, a procedure described in section 6.E.

3.3.4 Consistency with theoretical expectations

We assess the consistency of stated attendance with theoretical expectations by examining the coefficients estimated under models allowing separate coefficients for each attribute level, as well as comparing stated attendance and inferred attendance.

Recall that we expect that non-attendance to correspond to zero marginal utility. Thus, under the separate coefficients model, for positively preferred attributes, the coefficient on non-attended attributes should be lower than attended attributes. Conversely, for negatively preferred attributes, the coefficient on non-attended attributes should be higher. We may also expect to see coefficients of non-attended attributes to have higher variance, as in Campbell et al. (2008).

3.3.5 Consistency with inferred attendance

Stated attendance would also have greater face validity if it were consistent with inferred attendance, so we compared stated attendance using a polytomous attendance scale with inferred attendance. In this study we infer attendance in the manner of Hess and Hensher (2010), who used the respondent-specific conditional distribution of random coefficients to indicate attendance behaviour. The first step of this approach is to estimate a random parameters logit model (without any stated attendance data). The random parameters logit model is named because it is a generalization of the MNL model described above, in which the coefficients are drawn from a density for each individual, thus rendering them individual specific. Thus, the random utility is defined as

$$U_{ij} = X_{ij}\beta_i + \epsilon_{ij} \tag{3.4}$$

where β_i is the individual-specific draw of β from its distribution. Consequently, the choice probability expression for a random parameters logit model is

$$p_{ij} = \int \frac{\exp(V(X_{ij};\beta))}{\sum_m \exp(V(X_{im};\beta))} f(\beta) d\beta \tag{3.5}$$

where $f(\beta)$ is the pdf of β, an analyst-specified distribution. This generic distribution can be further refined by conditioning on observed choices because the observed choices reveal something about the individual's preferences; that is, the distribution of β describes the population of individuals, but once an individual's choices have been observed, we have some knowledge of where on the distribution they are located. We can thus compute a coefficient distribution conditional on an individual's responses. Train (2009) and Hess (2010) describe how to compute the mean and standard deviation of the respondent-specific conditional distribution.

$$\hat{\beta}_i = \frac{\sum_{r=1}^{R} p_i(y_i|\beta_r)\beta_r}{\sum_{r=1}^{R} p_i(y_i|\beta_r)} \tag{mean}$$

$$\tilde{\beta}_i = \sqrt{\frac{\sum_{r=1}^{R} \left(p_i(y_i|\beta_r)(\beta_r - \hat{\beta}_i)^2\right)}{\sum_{r=1}^{R} p_i(y_i|\beta_r)}} \tag{standard deviation}$$

where $\hat{\beta}_i$ is the estimated mean and $\tilde{\beta}_i$ the estimated standard deviation of the respondent-specific conditionl distribution, β_r is the rth draw from the estimated unconditional coefficient distribution, y_i is the observed choice sequence, and p_i is the probability of the observed choice sequence, given a specific value for β. Given a specific value of β, the choice probability of each choice occasion is independent from each other, and so the probability of observing a specific choice sequence, p_i, can can be computed as the product of the choice probability of the chosen alternative in each choice occasion.

$$p_i(y_i|\beta) = \prod_{k=1}^{K} \sum_{j=1}^{J} p_{ij}(\beta)^{y_{kj}} \tag{3.6}$$

where k indexes the choice occasion, j indexes the alternatives, $p_{ij}(\beta)$ is the choice probability in Equation 3.5 given a specific value for β, and y_{kj} is a dummy variable indicating whether the respondent chose alternative j on choice occasion k.

Once the respondent-specific conditional distributions have been derived, they can be used to infer attendance behaviour. Hess and Hensher (2010) argue for using a large coefficient of variation as an indicator of non-attendance, because this measure incorporates both the location and the uncertainty of the distributions. If we relied on the mean estimate alone, then near-zero estimates which indicate low, but non-zero preferences, may be incorrectly allocated to non-attendance. Hess and Hensher (2010) use a threshold of two for the coefficient of variation; cov values larger than two are assigned to attribute non-attendance. In this study, we follow Hess and Hensher (2010) and also use a threshold of two for the coefficient of variation when assigning dichotomous attendance/non-attendance states. We further extend the use of the coefficient of variation as an indicator for attendance state by treating it as an indicator for attendance *level*, analogous to the polytomous attendance scale used in the stated attendance part of the study. We compare the inferred attendance level to the stated attendance level (polytomous comparison) as well as the inferred attendance state to the stated attendance state (dichotomous comparison).

3.4 Data

The data for this study come from a choice experiment on preferences for designs of financial incentives in a minimal weight loss program performed via mail survey in late 2011 and early 2012. The study participants were recruited using the electronic patient database of Carilion Clinic, a healthcare organization in southwest Virginia. Obese and overweight patients were sought, and oversampling of males, minorities and low socioeconomic populations ensured a diverse sample.

The minimal weight loss program was described as being six months long, with the potential to lose one to two pounds per week. It included a number of behavioural components including a one-time dietitian meeting, weekly coaching calls, diet plan and tracking, exercise plan and tracking, and weekly weigh-ins at the program location. Figure 3.1 displays the components and their descriptions as presented on the survey instrument.

A literature review was conducted to form the initial set of attributes and attribute levels for the financial incentive designs. Payment frequency, reward magnitude, payment form and reward contingency were all attributes found to be varied in previous studies. From conducting focus groups on the survey instrument design, we also learned that program location had the potential to be an important attribute. Table 3.1 describes the final list of attributes and attribute levels used in our choice experiment, and Figure 3.2 displays the attributes and attribute levels as they were described in the survey instrument.

Choice tasks involved a choice set of two incentive designs and an opt-out alternative (see Figure 3.3 for an example). Choice sets were constructed using a D-efficiency design, in which a large number of possible designs were drawn from the full factorial design and sequentially compared with respect to the D-efficiency criterion[1]. A total of 96 choice

[1]The D-efficiency criterion refers to the goal of maximizing the determinant of the Fisher information

B

Weight control program

This section is about a weight control program to start people on the path to losing weight. It is 6 months long, which is 24 weeks. People who followed this program closely lost about

1-2 lbs	→	24-48 lbs
per week	results in	over 24 weeks

Each picture below represents one part of the program.

Dietitian meeting
Have a one-time meeting with a dietitian to develop personal diet and exercise plans.

Coaching calls
Receive weekly calls to help you follow your diet and exercise plans.

Diet plan
Use a personal workbook with recipes and more to plan healthy meals.

Diet tracking
Record meals and snacks online or on paper. Find out how many calories you ate.

Exercise plan
Use a personal workbook to plan exercise that's just right for you.

Exercise tracking
Record physical activity online or on paper. Find out how many calories you used.

Program weigh-ins
Weekly weight checks at program location. Only program staff will see your weight.

4

Figure 3.1: Description of behavioural weight loss intervention associated with financial incentive design, as presented to respondents.

Attribute	Attribute levels
Reward amount (reward)	\$0, \$48, \$96, \$216, \$384, \$576
Payment form (form)	Cash Pre-paid gym pass (gym) Health debit card for doctor's visits, prescriptions, and other medical expenses (medical) Debit card (debit)
Payment frequency (frequency)	Once at end of program (once) Quarterly Monthly Weekly
Program location (location)	Clinic Workplace Community center Church
Reward condition (condition)	Losing 2 lbs (weight) Attending weekly weight checks (attendance) Turning in records of diet and exercise (compliance) Attending weekly weight checks and turning in records of diet and exercise (att.comp)

Table 3.1: Attributes and attribute levels used in constructing the financial incentive. Abbreviated terms used in the text are in parentheses. Reward amount is defined as the total possible payment over the duration of the six-month intervention. Choice sets as presented to respondents included both the weekly reward amount and the total reward amount.

C Program Incentives

The weight control program described in Section B can include incentives for participation. The information on this page is about different incentives that can be added to the program.

Weigh in weekly

Program location

This is where you go for the weekly weight checks and meet with the dietitian. It is *not* where you have to exercise.

Examples
- Clinic
- Workplace
- Community center
- Church

Achieve weight control goal

Weekly weight control goal

You must achieve this goal to earn the reward each week. Some weeks you may not meet the goal. In those weeks you will not earn the reward.

Examples
- Losing 2 lbs
- Turning in complete records of diet and exercise
- Attending weekly weigh-ins

Earn weekly reward

Weekly reward available

This is the dollar value of the weekly reward. You have an opportunity each week to earn this reward by achieving your weekly weight control goal.

Examples
- $5 per week
- $24 per week
- $36 per week

Get paid periodically

Payment frequency

This is how often you get paid the rewards you have earned. Remember that you earn rewards by achieving the weekly weight control goal.

Examples
- Weekly
- Monthly
- Quarterly
- Once at program end

Receive payment

Payment form

This is how you are paid. Each type will have the same dollar value as if you were paid in cash.

Examples
- Cash or debit card
- Pre-paid gym pass
- Health debit card for doctor's vists, prescriptions, and other medical expenses

6

Figure 3.2: Description of financial incentive attributes and attribute levels as presented in the survey instrument.

sets were constructed, and each survey instrument contained four choice sets, resulting in 24 instrument versions. These were assigned randomly and distributed as evenly as possible to the survey sample.

After the choice tasks, attribute attendance was elicited using a 6-point Likert scale, from 1 for 'Never' to 6 for 'Always' (Figure 3.4). Note that attribute attendance was sought for 'weekly reward' and 'total reward' separately, even though the total reward was computed from the weekly reward. We chose this format to be consistent with the choice tasks, in which both the 'weekly reward' and 'total reward' were presented for the convenience of the respondent. Half of the survey sample was randomly assigned to receive the attribute attendance question; the other half received a different question which is not the subject of this study.

3.5 Results

A total of 7,554 individuals were selected for recruitment using the Carilion Clinic electronic patient database. Of those recruited, 2,737 individuals consented to the survey during phone recruitment. Failure to recruit at this stage was primarily due to the high rate (59%) of invalid phone numbers in the electronic patient database. Of those individuals who were reachable by phone, 88% consented to the survey. One all follow-ups were completed, 1,297 completed surveys were received (47% survey completion rate). Of the completed surveys, one survey was excluded because the respondent was younger than 18 years old. Hence, the final number of surveys analyzed was 1,296. Of those 1,296 completed surveys, 671 (52%) included the post-choice task attribute attendance question.

Figure 3.5 displays the distribution of attendance levels by individual attribute, sorted by the proportion of high responses (4, 5 or 6). The proportion of low responses (1, 2 or 3) is substantial, ranging from 20% to 50% depending on the attribute. The number of 1s (labelled 'never' considered) often outnumbers either of the other low responses. These rates of stated non-attendance are not unusual and well within the broad range of non-attendance rates found in previous studies (Alemu et al., 2013).

3.5.1 Baseline

We first present the baseline model, which is a multinomial logit model assuming full attendance to all attributes; that is, no attendance information is used (Table 3.3). For each

matrix of the parameter estimates

1 Please consider the following two weight control programs.

	Program A	Program B
Program location	Clinic	Church
Weekly weight control goal	Attending weekly weight checks	Turning in records of diet and exercise
Weekly reward available	$16	$4
Payment frequency	Once at end of program	Weekly
Payment form	Health debit card for doctor's visits, prescriptions, and other medical expenses	Debit card
Total reward available in program	$16 × 24 wks = $384	$4 × 24 wks = $96

Which weight control program would you choose?
Please check one box.

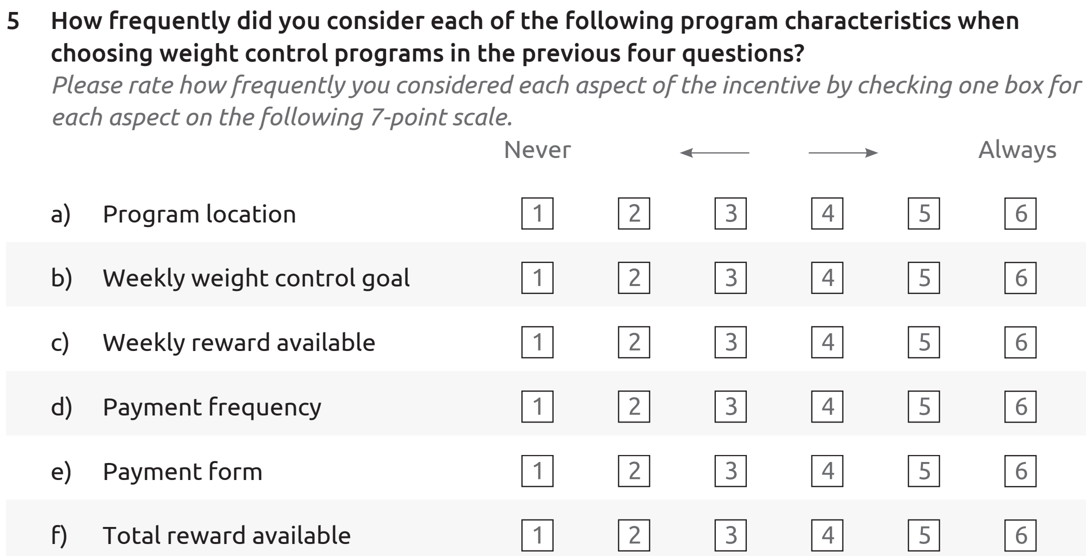
☐ 1 Program A

☐ 2 Program B

☐ 3 I would not choose either program.

114 10

Figure 3.3: Example of choice question



5 How frequently did you consider each of the following program characteristics when choosing weight control programs in the previous four questions?

Please rate how frequently you considered each aspect of the incentive by checking one box for each aspect on the following 7-point scale.

		Never		←	→		Always
a)	Program location	1	2	3	4	5	6
b)	Weekly weight control goal	1	2	3	4	5	6
c)	Weekly reward available	1	2	3	4	5	6
d)	Payment frequency	1	2	3	4	5	6
e)	Payment form	1	2	3	4	5	6
f)	Total reward available	1	2	3	4	5	6

Figure 3.4: Attribute attendance question, presented after choice tasks.

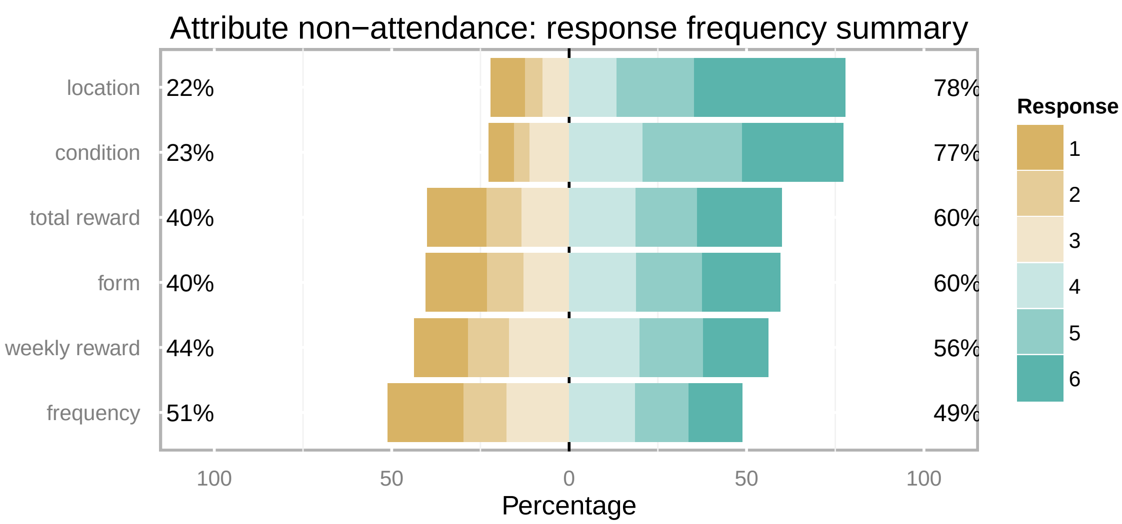


Figure 3.5: Distribution of responses to pre-choice task attribute importance question. The percentages on the right indicate the proportion of responses which are high (4, 5 or 6, where 6 indicates 'always' attended), while the percentages on the left indicate the proportion of low responses (1, 2 or 3, where 1 indicates 'never' attended).

attribute, at least one level is significantly different from zero, with the location attribute containing three levels that have significant coefficients. The clinic and community center locations are preferred to the workplace. For the payment form attribute, debit card is preferred above other levels, just as weekly payments is preferred to other payment frequencies. On the other hand, requiring both attendance and record-keeping is preferred less than other reward conditions. As expected, the reward amount is highly significant, although it was one of the attributes with the highest proportion of low responses to the attendance question.

3.5.2 Non-attendance as zero marginal utility

We next consider which points on the attendance scale are most consistent with a theoretical expectation of zero marginal utility. We do so by estimating a set of models where coefficients on ignored attributes are constrained to zero at the respondent level, and different segments of the attendance scale indicate attribute non-attendance. Each model corresponds to grouping adjacent points on the attendance scale into the 'ignored' part of the scale, and the list of models represents an exhaustive combination of such groupings. Since each model has the same number of estimated parameters and the same number of observations, we compare them using log-likelihood values (Figure 3.6). The models with the highest log-likelihood values are those which assign the negative responses (points 1, 2 and 3) to non-attendance, with the lowest response (point 1) garnering the maximum log-likelihood values in this set. This result is reassuring, because it displays consistency between our expectations given the wording of the attendance question (where 'never' is associated with point 1) and the empirical result (where point 1 best corresponds to zero marginal utility).

However, comparing log-likelihood values only allows us to rank models without the ability to say whether one model is significantly better than another. Thus, we refine our empirical results by employing the Vuong test for non-nested hypotheses (Vuong, 1989) in order to discriminate between the top three models, as ranked by log-likelihood values (Table 3.2). The results show that the Vuong test is unable to discriminate between the top two models, so we cannot reject the null that the two models are equally distant from the true specification. However, the test does discriminate between the second and third models, in favour of the second model. Thus, it is clear that the top two models are significantly different from the remaining models in this set, and therefore, points 1 and 2 on the attendance scale are the best indicators of non-attendance. However, whether it is point 1 alone or also point 2 which are the best indicators is unclear. These results are similar to the practice in Scarpa et al. (2013) of assigning the two lowest points to non-attendance and the finding in Colombo et al. (2013) that the lowest point corresponds to non-attendance, but is not consistent with the decision in Weller et al. (2014) to assign

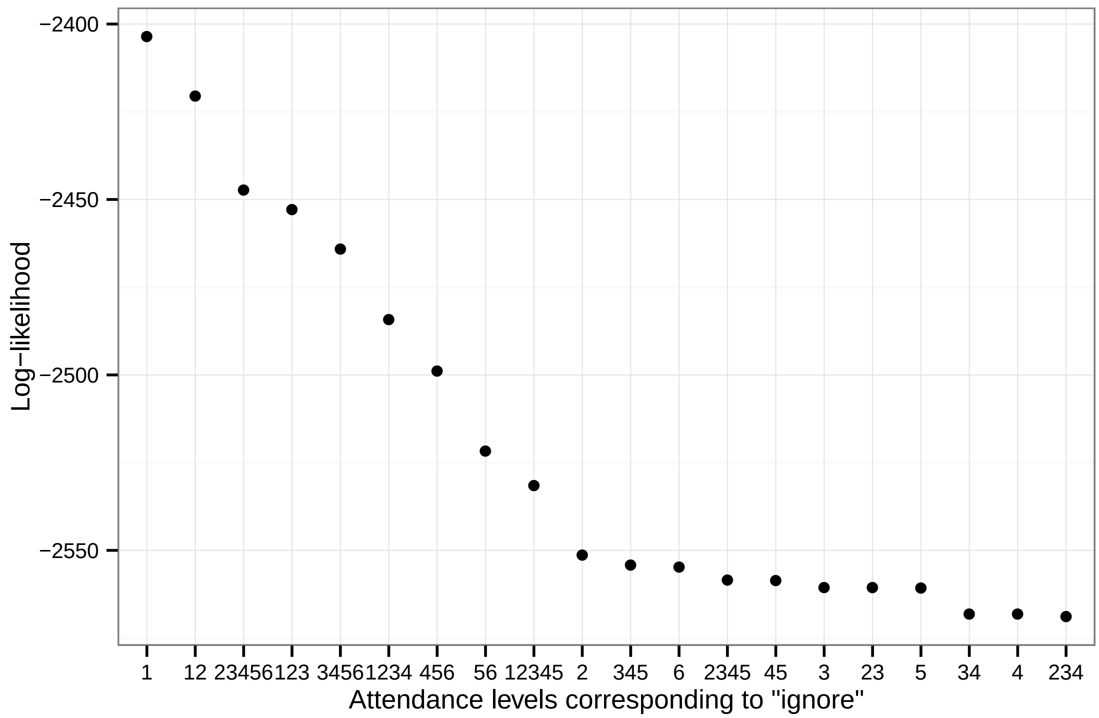


Figure 3.6: Log-likelihood values of models where coefficients on ignored attributes are constrained to zero at the respondent level.

the three lowest points, including the middle point, to non-attendance.

Model 1	Model 2	Test statistic	Decision
1 as ignore	12 as ignore	1.43	Unable to discriminate
12 as ignore	123 as ignore	2.78	Reject in favour of Model 1

Table 3.2: Vuong test results for top-ranked models where coefficients on ignored attributes are constrained to zero at the respondent level. The test statistic is distributed as a standard normal, and the decision rule uses a critical value of 1.96.

Comparing the favoured models in this set to the baseline model, we see a remarkable improvement in model fit according to all measures (Table 3.3). This result is consistent with other studies which have taken into account attribute non-attendance, the vast majority of which have found improved fit. The sign of significant coefficients is similar across all models, although there are a few changes in significance. For example, constraining coefficients on ignored attributes to zero seems to have sharpened preference estimates for payment form, revealing a negative preference for the health debit card compared to other levels. In general, the changes in significance are due to modest changes in coefficient estimates and standard errors, reflecting the marginal significance level of the coefficients rather than any dramatic change.

Table 3.3: (continued on next page)

	Baseline	1 as ignore	12 as ignore
ASC.SQ	-0.290^{**}	-0.409^{***}	-0.210^{***}
	(0.104)	(0.048)	(0.039)
Mag.0	0.027	-0.112	0.081
	(0.109)	(0.065)	(0.058)
log(amount + 1)	0.236^{***}	0.358^{***}	0.305^{***}
	(0.039)	(0.021)	(0.018)
location: workplace	-0.163^{***}	-0.200^{***}	-0.197^{***}
	(0.049)	(0.053)	(0.054)
location: community	0.109^{*}	0.130^{*}	0.121^{*}
	(0.049)	(0.052)	(0.054)
location: church	-0.066	-0.031	-0.017
	(0.048)	(0.051)	(0.052)
location: clinic	0.120^{*}	0.101^{*}	0.093
	(0.048)	(0.051)	(0.053)
form: gym	-0.036	-0.082	-0.098
	(0.052)	(0.058)	(0.062)
form: medical	-0.080	-0.123^{*}	-0.135^{*}
	(0.051)	(0.055)	(0.059)
form: debit	0.129^{*}	0.143^{*}	0.173^{**}
	(0.051)	(0.056)	(0.060)
form: cash	-0.013	0.061	0.059
	(0.056)	(0.059)	(0.063)
condition: weight	0.083	0.078	0.067
	(0.052)	(0.055)	(0.056)
condition: compliance	-0.040	-0.062	-0.082
	(0.052)	(0.055)	(0.056)
condition: att.comp	-0.108^{*}	-0.109^{*}	-0.103
	(0.050)	(0.053)	(0.054)
condition: attendance	0.065	0.093	0.119^{*}
	(0.054)	(0.057)	(0.058)

Table 3.3: (continued from previous page)

	Baseline	1 as ignore	12 as ignore
frequency: weekly	0.111*	0.109	0.146*
	(0.055)	(0.062)	(0.068)
frequency: monthly	−0.081	−0.014	−0.036
	(0.052)	(0.059)	(0.064)
frequency: quarterly	0.059	0.017	−0.014
	(0.050)	(0.056)	(0.062)
frequency: once	−0.089	−0.113	−0.096
	(0.052)	(0.059)	(0.063)
Log-likelihood	−2543.513	−2403.606	−2420.547
AIC	5117.025	4837.212	4871.094
BIC	5204.398	4924.584	4958.467
$\bar{\rho}^2$	0.049	0.101	0.094
N	2502.000	2502.000	2502.000

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.3: Coefficient estimates for baseline model and top-ranked models assuming non-attendance as zero marginal utility.

3.5.3 Attendance scale length

Now we consider the value of different attendance scale lengths by estimating models under recoded attendance scales, constructed by collapsing adjacent points on the original six-point scale into common attendance levels. These models are specified with separate coefficients for each attendance level. We evaluate these models according to a number of model selection criteria, including prediction accuracy, AIC, BIC, and McFadden's pseudo R^2, adjusted for degrees of freedom (Ben-Akiva and Lerman, 1985).

Figure 3.7 presents models estimated under all possible recoded attendance scales, ordered by increasing prediction error. From the figure, we can see that the lowest error model is the one estimated under a recoded dichotomous scale, where points 1 through 5 are assigned to one level and point 6 is assigned to another level. Judging by the 95% coverage interval, this model has significantly less error than the next lowest error model. Hence we choose this model as a candidate for further analysis via hypothesis

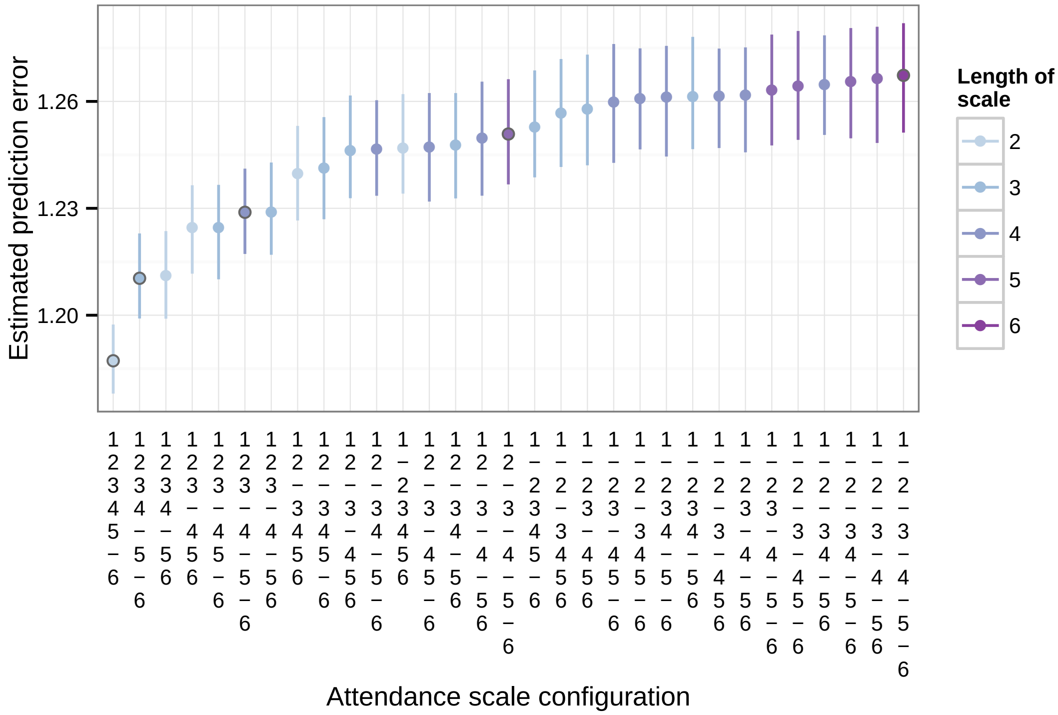


Figure 3.7: Prediction accuracy, as measured by the multinomial deviance loss function, for models estimated under all possible recoded attendance scales. Error bars represent 95% coverage intervals. The minimum error model for each attendance scale length is highlighted with a grey outline.

testing. In addition to configuration 12345|6, other candidates suggested by the figure include 1234|5|6 and 123|4|5|6, because they are the minimum error model for their given attendance scale lengths and because they may have significantly different error from other models.

Figure 3.8 presents the same models in the same order as Figure 3.7, but the measures presented are goodness-of-fit measures instead. A striking difference between the two figures is that prediction accuracy clearly favours very different models from goodness-of-fit measures. Both AIC and $\bar{\rho}^2$ favour longer attendance scales, thereby ranking the models in almost the opposite order from prediction accuracy. BIC favours parsimony more than the other goodness-of-fit measures, and its rankings appear to be driven primarily by attendance scale length. According to the goodness-of-fit measures, there are two natural candidates for further for hypothesis testing: 1|23456 and 1|234|56. These models not only have some of the highest rankings in each measure, they have very similar rankings across all three measures.

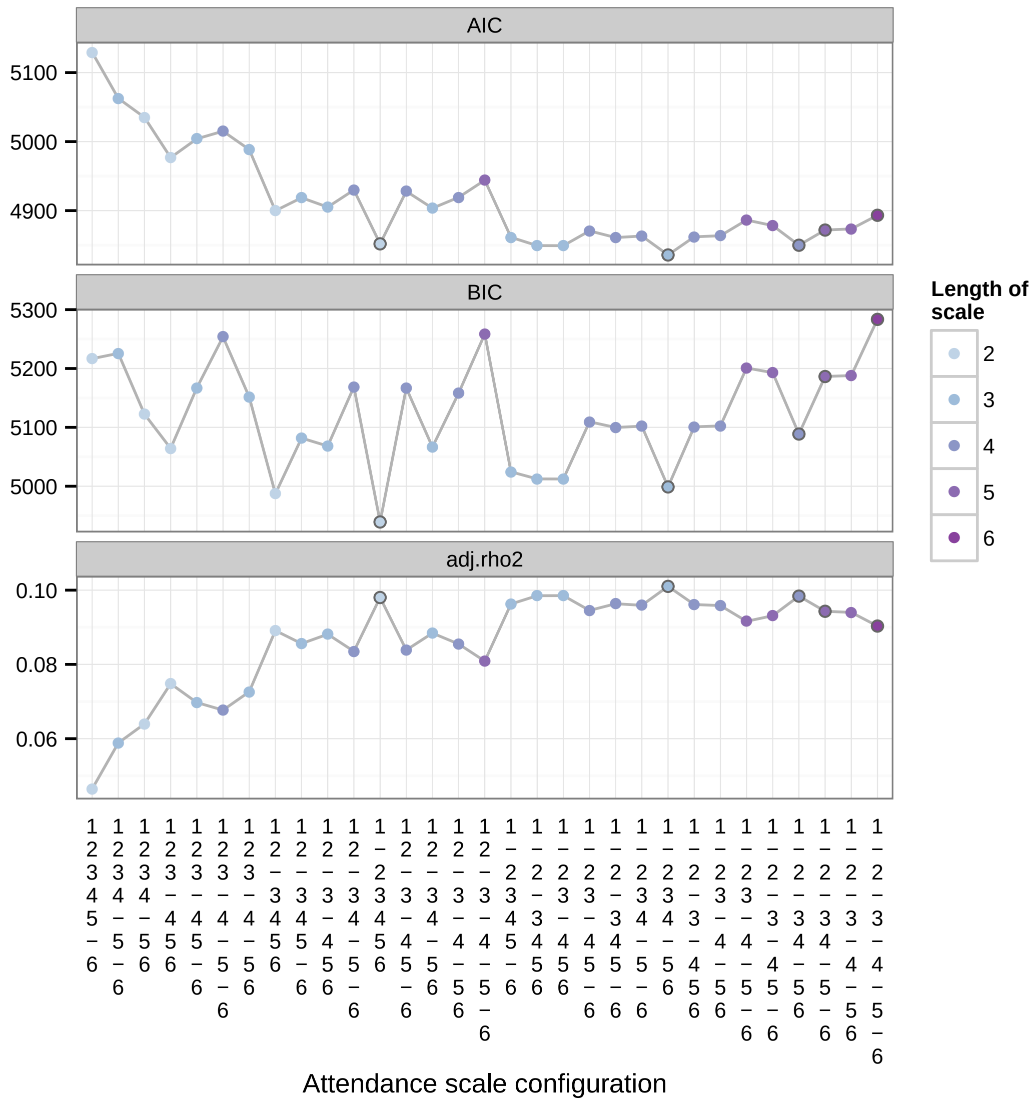


Figure 3.8: Goodness-of-fit measures, including AIC, BIC and McFadden's pseudo R^2 (adjusted for degrees of freedom), for models estimated under all possible recoded attendance scales. The models are in the same order as in Figure 3.7. The best model for each attendance scale length, according to a given measure, is highlighted with a grey outline.

We now have two sets of candidate models to compare against the original six-point attendance scale, one set determined by prediction accuracy and the other set determined by goodness-of-fit measures. Table 3.4 presents the results of likelihood ratio tests of each candidate model against the model estimated under the original six-point attendance scale. For each candidate model selected by prediction accuracy, we can reject the null that the candidate model and the fully polytomous model are the same at the 95% significance level. However, for each candidate model selected by goodness-of-fit measures, we *cannot* reject the null that the candidate model and the fully polytomous model are the same.

Attendance scale	DoF	Log-likelihood	χ^2	p-value
1 | 2 | 3 | 4 | 5 | 6	67	−2379.6		
12345 | 6	15	−2549.7	340.1	0.000
1234 | 5 | 6	28	−2503.3	247.4	0.000
123 | 4 | 5 | 6	41	−2466.5	173.8	0.000
1 | 23456	15	−2410.9	62.6	0.148
1 | 234 | 56	28	−2389.9	20.5	0.993

Table 3.4: Likelihood ratio tests of candidate models against model estimated under original six-point attendance scale. The first row indicates the original six-point scale, the second group of rows the candidate models selected by prediction accuracy, and the third group of rows the candidate models selected by goodness-of-fit measures.

The opposing results from the two sets of candidate models give opposite interpretations: the models selected by prediction accuracy imply that the original, longer scale is more informative than the shorter scale, whereas the models selected by goodness-of-fit measures imply that the original, longer scale is *not* more informative than the shorter scale. To resolve this contradiction, we test candidate models from one set against the other set using the Vuong test for non-nested hypotheses.

We select one candidate model from each set for the Vuong test, 12345 | 6 and 1 | 23456, chosen because they each ranked best in their respective sets and because they are the most parsimonious within each set. The test statistic for the two models chosen is -7.18. Using a significance level of 95%, which corresponds to a critical value of 1.96, the test indicates that the null hypothesis of equally distant specifications can be rejected in favour of the second model, 1 | 23456.

Hence, the test results suggest relying on the interpretation arising from the set of candidate models selected by goodness-of-fit measures: that the original six-point scale is not more informative than the shorter dichotomous and trichotomous scales which were recoded. However, a likelihood ratio test of the dichotomous scale (1 | 23456) against the trichotomous scale (1 | 234 | 56) reveals a significant difference (χ^2 : 42.1). Thus, a trichotomous scale may be more informative than a dichotomous scale, similar to the finding in Colombo et al. (2013).

3.5.4 Consistency with theoretical expectations

Intuition and the results in Colombo et al. (2013) suggest that the attendance levels defined in the two candidate models could correspond to non-attendance, partial attendance and full attendance. We attempt to verify these interpretations by examining the estimated coefficients for both candidate models. Due to the number of estimates involved, we consider a graphical presentation of them in the text, while making a tabular form available in section 6.F. Figure 3.9 displays the coefficient estimates for the model estimated under the original six-point attendance scale, while Figure 3.10 and Figure 3.11 display the coefficient estimates for the models estimated under the favoured dichotomous and trichotomous models, respectively. In each figure, the effects and their 90% confidence intervals are plotted as point ranges, with effects significantly different from zero in black and insignificant effects in grey. In the background, the effects and their 90% confidence intervals for the full attendance model are also plotted in blue.

Under the original six-point scale, the coefficients on the same attribute level across different attendance levels are, for the most part, not significantly different from each other, suggesting either insufficient sample sizes to pick up significant effects or that attendance levels may not be associated with different preference intensities. A notable exception is the coefficients associated with reward amount, which display a very clear upward trend, indicating that the greater the stated attendance to reward amount, the more positive and intense the sensitivity to reward amount. For this attribute alone, the results suggest some conflation of preference and stated attendance.

Under the dichotomous scale model, about half of the effects are significantly different between the two attendance groups, providing limited evidence for consistency with theoretical expectations. Furthermore, the 'ignored' attendance group frequently has a more positive effect than the 'attended' group, even for preferred attribute levels such as debit card. This observation is at odds with the theoretical expectation that the 'ignored' attendance group should have a more negative effect than the 'attended' group for preferred attribute levels.

Under the trichotomous scale model, the effects tend to increase or decrease monotonically across attendance levels for the same attribute level. This pattern supports interpretation of the points on the trichotomous scale as moving from less attendance ('no attendance') to more attendance ('partial' and then 'full attendance'), as in Colombo

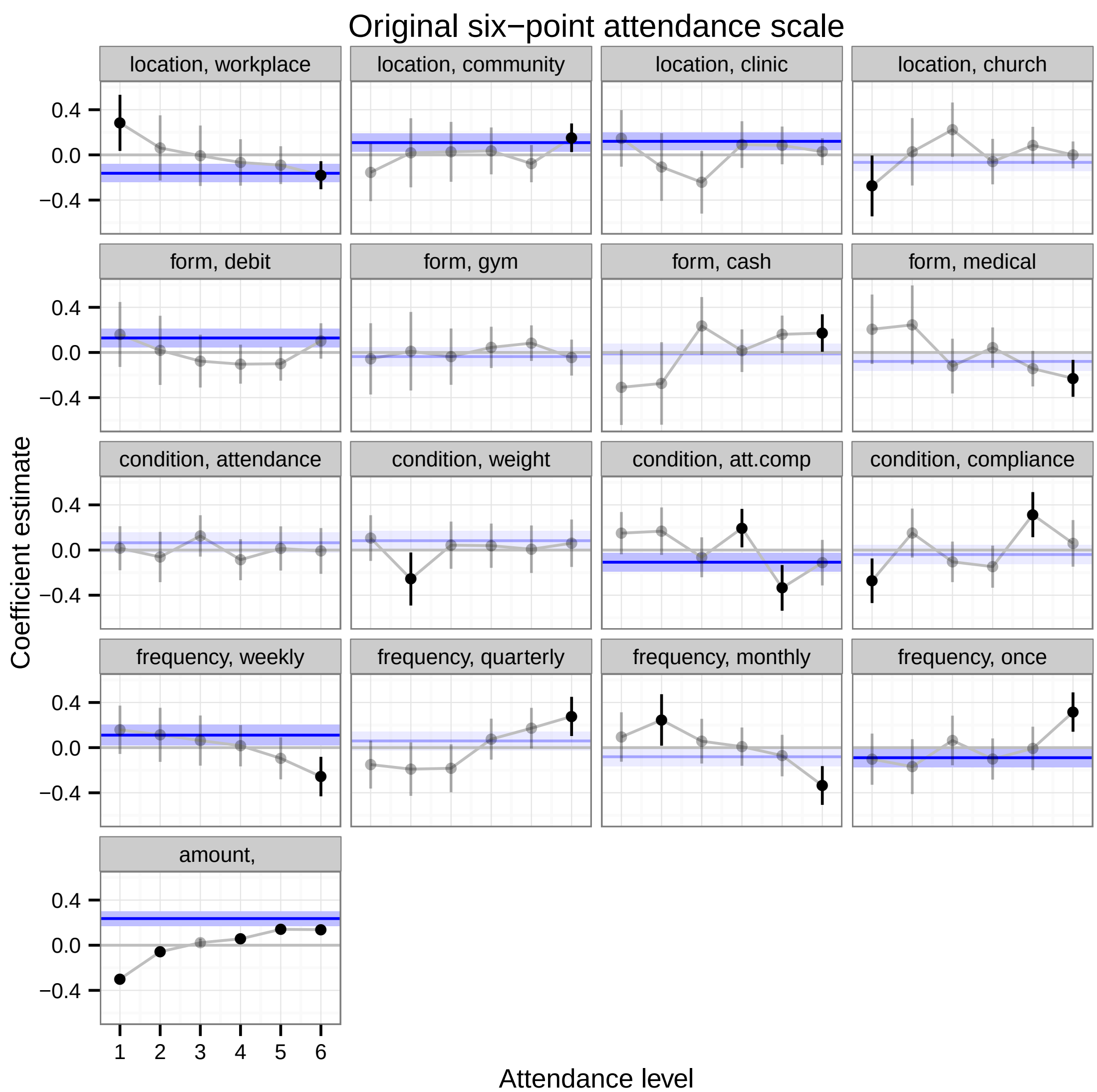


Figure 3.9: Coefficients and 90% confidence intervals by attendance group and attribute level estimated under a model with separate coefficients for each point on the original six-point attendance scale. The blue line and ribbon in the background represent the full attendance model coefficients and 90% confidence intervals, respectively. Significant effects are darker than insignificant effects.

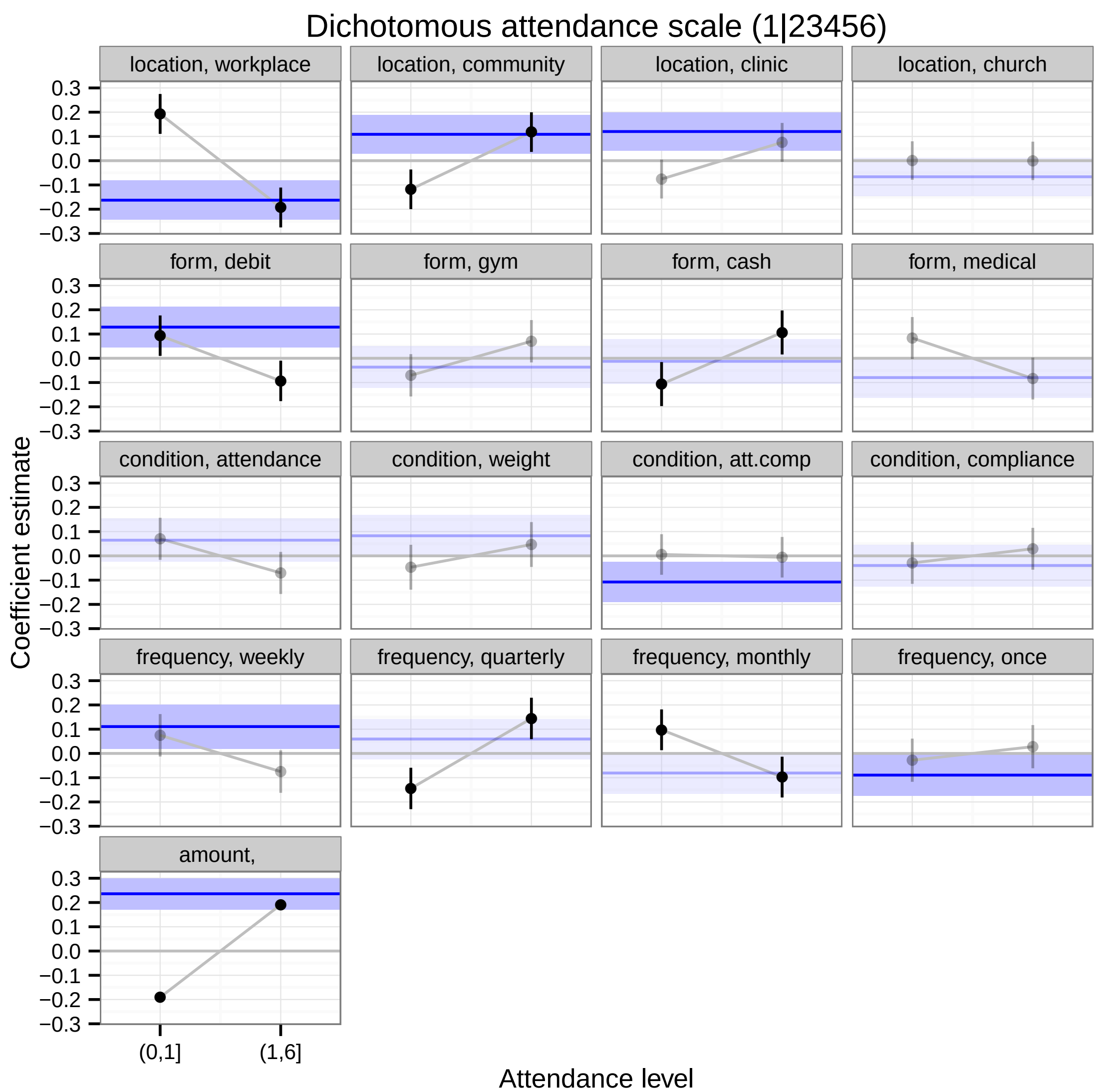


Figure 3.10: Coefficients and 90% confidence intervals by attendance group and attribute level estimated under a model with separate coefficients for each level of a recoded dichotomous attendance scale. The blue line and ribbon in the background represent the full attendance model coefficients and 90% confidence intervals, respectively. Significant effects are darker than insignificant effects.

et al. (2013). However, the direction of the trend does not always follow theoretical expectations. Similar to the dichotomous scale model, lower attendance levels are sometimes associated with more positive effects even for preferred attribute levels such as debit card, casting serious doubts on the interpretation of the trichotomous scale. We may also expect that the variance of ignored attribute levels to be greater, as in Campbell et al. (2008), but again, we do not find this pattern to be the case consistently.

3.5.5 Consistency with inferred attendance

We next consider consistency between stated and inferred attendance, where attendance is inferred using the respondent-specific conditional distribution. The conditional distribution is computed by first estimating a random parameters logit model with a similar specification to the baseline model of subsection 3.3.1, except that the coefficients on each attribute are specified to have a normal distribution. Second, the mean and standard deviation of the conditional distributions are computed using 500 Halton sequence draws from the estimated normal distribution.

In Figure 3.12, we consider a continuous level of inferred attendance. In this case, inferred attendance for a specific individual and attribute is based on the coefficient of variation of the conditional distribution of the random coefficient corresponding to that attribute for that individual. The higher the coefficient of variation, the higher the inferred non-attendance. If inferred attendance and stated attendance corresponded well to each other, we would expect to see that higher stated attendance corresponded to lower inferred non-attendance. However, few random coefficients display this pattern, the exceptions being the coefficients on reward amount and the debit card payment form.

We bolster visual inspection with a statistical model, estimating an ordered logistic model where the stated attendance level for a specific attribute is the dependent variable and the inferred attendance levels (coefficient of variation values) for each level of that specific attribute are the explanatory variables (Table 3.5). If inferred attendance and stated attendance corresponded well to each other, we would expect to see significant negative coefficients in the ordered logistic model. Instead, we see mostly insignificant coefficients, with only one significant negative coefficient (on the debit card payment form, which we identified visually as displaying the expected pattern) and, unexpectedly, a significant positive coefficient.

In Figure 3.13, we consider dichotomous stated and inferred attendance, using the lowest point as non-attendance for stated attendance and a threshold of 2 for the coefficient of variation as Hess and Hensher (2010) does for inferred attendance. For every attribute or attribute level, stated and inferred sources disagree on non-attendance more than they agree. That is, the number of dissimilar classifications are far greater than the

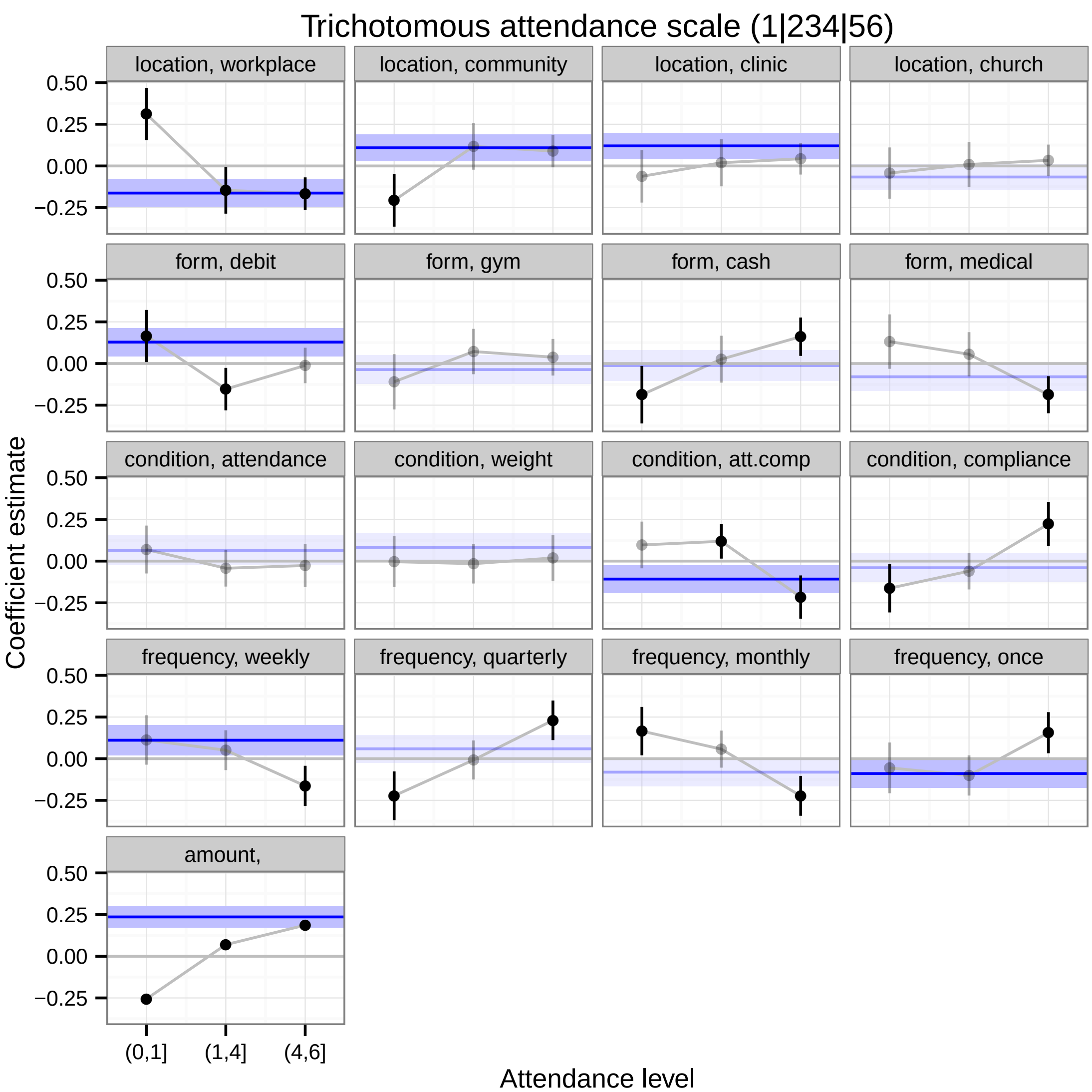


Figure 3.11: Coefficients and 90% confidence intervals by attendance group and attribute level estimated under a model with separate coefficients for each level of a recoded trichotomous attendance scale. The blue line and ribbon in the background represent the full attendance model coefficients and 90% confidence intervals, respectively. Significant effects are darker than insignificant effects.

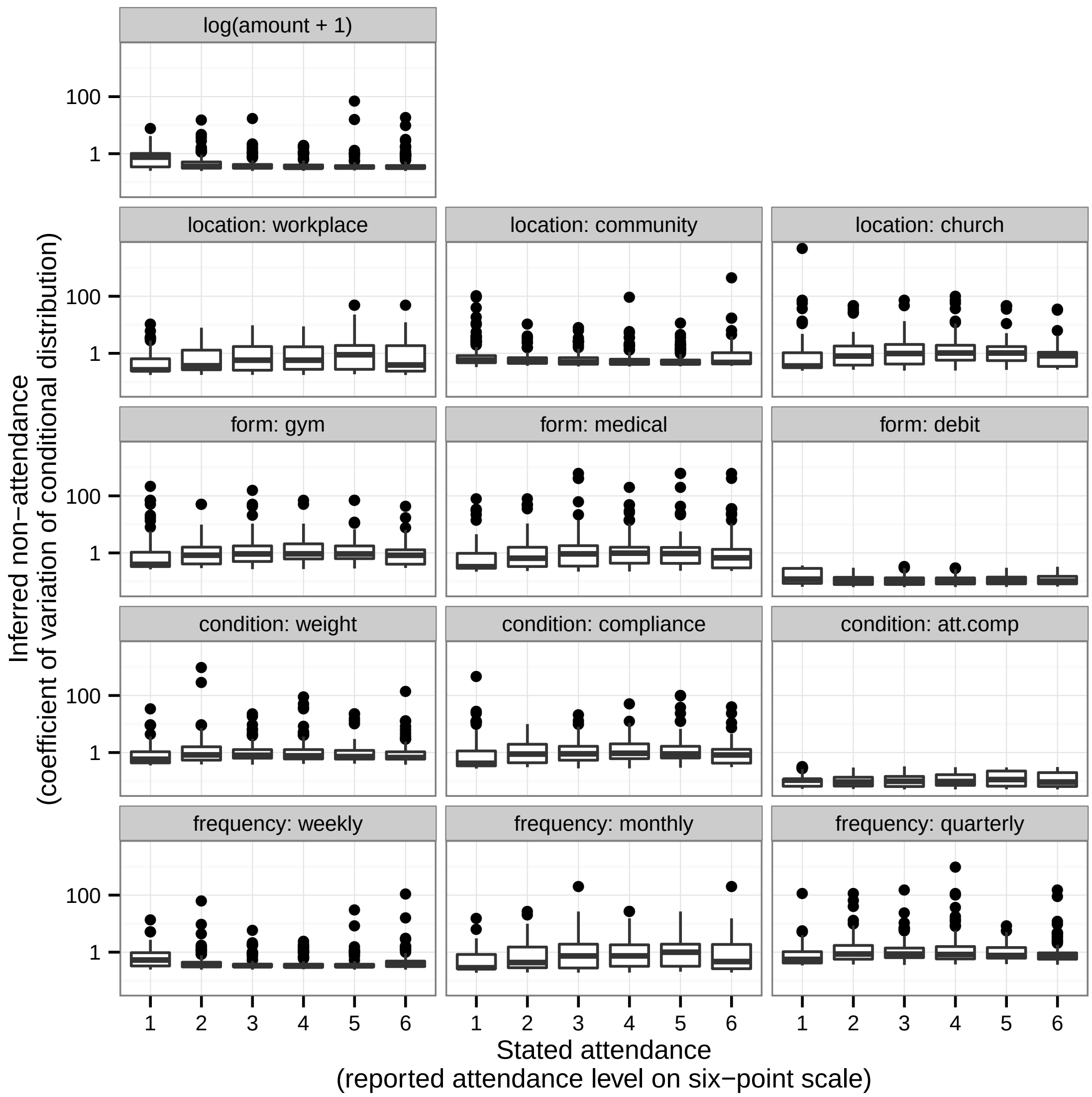


Figure 3.12: Box plot of the distribution of inferred attendance level (measured by the respondent-specific conditional distribution coefficients of variation) over individuals by stated attendance level and random coefficient.

	amount	location	form	condition	frequency
log(amount + 1)	-0.01				
	(0.02)				
location: workplace		0.02			
		(0.02)			
location: community		0.02			
		(0.01)			
location: church		0.00			
		(0.00)			
form: gym			-0.00		
			(0.00)		
form: medical			0.00		
			(0.00)		
form: debit			-8.83^{***}		
			(1.18)		
condition: weight				0.00	
				(0.00)	
condition: compliance				0.03^{*}	
				(0.02)	
condition: att.comp				1.30	
				(0.95)	
frequency: weekly					0.01
					(0.02)
frequency: monthly					0.01
					(0.01)
frequency: quarterly					0.00
					(0.00)
AIC	2242.67	1960.65	2196.21	2048.81	2268.06
BIC	2269.39	1996.28	2231.84	2084.44	2303.69
Log Likelihood	-1115.33	-972.33	-1090.11	-1016.40	-1126.03
Deviance	2230.67	1944.65	2180.21	2032.81	2252.06
Num. obs.	635	635	635	635	635

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.5: Coefficient estimates for ordered logistic regression of stated attendance on inferred attendance (measured by respondent-specific conditional distribution coefficient of variation values).

number of cases classified as non-attendance by both stated and inferred sources. For some attributes and attribute levels, there are far more cases of 'silent attend' than 'silent ignore'; that is, respondents tend to state non-attendance when attendance is inferred. The disproportionate number of 'silent attend' cases was found by previous studies (Hess and Hensher, 2010; Collins et al., 2013; Campbell et al., 2011), and one reason why self-reported attendance behaviour was questioned. On the other hand, several attribute levels have nearly equal numbers of 'silent attend' cases and 'silent ignore' cases, which indicates that neither source of attendance information is systematically biased with respect to the other.

3.6 Conclusion

Attribute non-attendance has attracted a growing amount of attention in the last decade, but researchers still struggle with how to interpret non-attendance statements. The reliability of self-reported attendance is suspect due to the potential for misreporting, conflation between low and zero preferences, and the potential for attendance to differ depending on the choice set presented. Polytomous attendance scales have the potential to ameliorate all three of these effects, by potentially decreasing social desirability bias, differentiating between low and zero preferences, and providing an avenue for respondents to indicate different frequencies of attribute attendance. However, the increased respondent burden and model complexity calls for a closer look at these potential benefits.

This study uses a choice experiment in health economics to assess the performance and theoretical face validity of a six-point polytomous attendance scale. The first part of the analysis considered which points on the six-point attendance scale was most consistent with the theoretical interpretation of non-attendance as zero marginal utility. Using log-likelihood values and the Vuong test for non-nested hypotheses, we found that the lowest point or the lowest two points were the best indicators of non-attendance, interpreted as zero marginal utility. This result is consistent with the finding in Colombo et al. (2013) that the lowest point corresponds to non-attendance, and the practice in Scarpa et al. (2013) of assigning the two lowest points to non-attendance. However, this result is not consistent with the decision in Weller et al. (2014) to assign the lowest three points, including the midpoint, to non-attendance. Similar to the vast majority of other attribute non-attendance studies, imposing zero marginal utility on non-attended attributes dramatically improved model fit.

The second part of the analysis assessed the value of longer attendance scale by comparing the original six-point attendance scale with scales recoded to a shorter length. In this part of the analysis, we recoded the original attendance scale by grouping adjacent

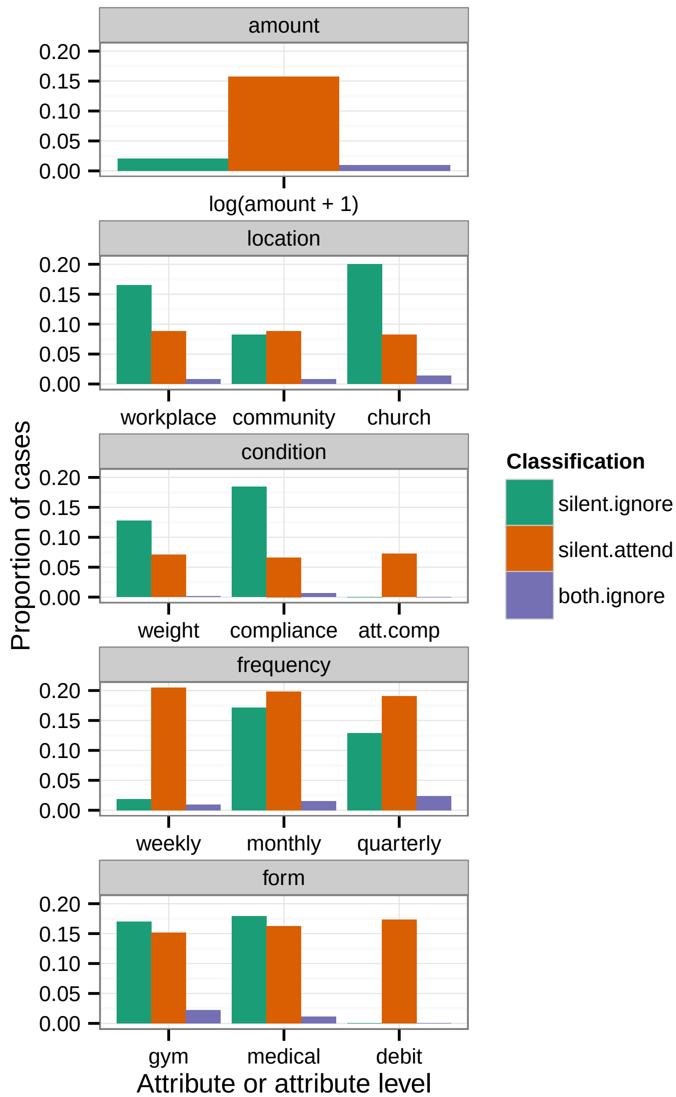


Figure 3.13: Histogram of dissimilarities and similarities between stated attendance and inferred attendance. Stated non-attendance occurs when 'Silent ignore' refers to stated attendance and inferred non-attendance. 'Silent attend' refers to stated non-attendance and inferred attendance. 'Both ignore' refer to cases where non-attendance is both stated and inferred.

points into new attendance levels, thus forming artificially shorter scales. We then estimated MNL models with separate coefficients for each attendance level, and selected the best models using criteria such as prediction accuracy and goodness-of-fit. After using the Vuong test for non-nested hypotheses to select between the candidate models selected by different criteria, we found that the favoured models were estimated under dichotomous and trichotomous models which assigned the lowest point on the six-point scale to its own attendance level. The trichotomous scale further split the higher points into two attendance levels (234 and 56).

The third part of the analysis compared empirical findings with theoretical expectations. While we expected the attendance levels in the dichotomous scale to correspond to non-attendance and attendance, and the attendance levels in the trichotomous scale to correspond to non-attendance, partial attendance and full attendance, close examination of the coefficient estimates under the dichotomous and trichotomous scales cast doubt on these interpretations. There were several instances of inconsistency with theoretical expectation, such as coefficients associated with 'ignored' attributes being more positive than coefficients associated with attended attributes.

Finally, we compared stated attendance with inferred attendance, where attendance was inferred from the coefficient of variation of the respondent-specific conditional distribution on the random parameters of a random parameters logit model. We compared a polytomous version of stated attendance to a continuous version of inferred attendance, and found, both visually and statistically, that higher stated attendance did not correspond to lower inferred non-attendance. We also compared a dichotomous versions of stated and inferred attendance to each other, and found more dissimilar non-attendance classifications than similar.

These results suggest two conclusions: 1) that a six-point scale is unnecessarily long, and a dichotomous or trichotomous scale, as found previously in the literature, is likely to be sufficient to capture what attendance information there is, and 2) that the nature of the attendance information captured is still enigmatic to the choice modeller. Although imposing zero marginal utility on non-attended attributes improves model fit, freeing the model from this theoretical constraint presents a very different picture, in which 'non-attended' attributes often display non-zero marginal utility. Furthermore, stated attendance reported on a polytomous attendance scale was no more consistent with inferred attendance than previous studies have found.

Thus, the search for more reliable attendance elicitation formats continues. As mentioned in the literature review, collecting attendance state through implicit measures, as suggested by Hensher et al. (2005), seems to be one potential way forward. Another possible direction is to combine stated and inferred attendance, by treating attendance statements as measurements with error, and allowing them to inform but not dictate attendance state. Hensher et al. (2007) made an early attempt at treating knowledge of the attendance state stochastically, by modelling attendance behaviour as a type of 'choice' based on expected utility maximization. The expected maximum utility index derived from estimates for the attendance 'choice' model is then used as a covariate in the usual choice model representing choices made in the choice experiment. This novel specification was found to outperform the conventional model assuming full attendance in terms of model fit. A later attempt was the hybrid choice model developed by Hess and Hensher (2013) and mentioned earlier in the literature review. Positing a latent variable which drove both attendance statements and choice, the hybrid model addressed not only the possibility of measurement error in attendance statements, but also endogeneity concerns. Since both attendance statements and choices were dependent variables in the hybrid model, the potential endogeneity between them was explicitly modelled. Empirically, the study found that the probability of stating non-attendance decreases with the latent variable, which supports its interpretation as attribute importance. However, the hybrid model did not outperform the full attendance model. Collins et al. (2013) and Hole et al. (2013) took an approach which is broadly applicable to any latent class approach: they allowed attendance statements to enter as covariates which help explain class membership probabilities. Both studies found that including attendance statements in this manner improved model fit greatly.

Bibliography

Alemu, Mohammed Hussen, Mørkbak, Morten Raun, Olsen, Søren Bøye, and Jensen, Carsten Lynge. Attending to the reasons for attribute non-attendance in choice experiments. *Environmental and Resource Economics*, 54(3):333–359, 2013. ISSN 0924-6460. doi: 10.1007/s10640-012-9597-8. URL http://dx.doi.org/10.1007/s10640-012-9597-8.

Balcombe, Kelvin, Burton, Michael, and Rigby, Dan. Skew and attribute non-attendance within the bayesian mixed logit model. *Journal of Environmental Economics and Management*, 62(3):446 – 461, 2011. ISSN 0095-0696. doi: http://dx.doi.org/10.1016/j.jeem.2011.04.004. URL http://www.sciencedirect.com/science/article/pii/S0095069611000635.

Balcombe, Kelvin, Bitzios, Michail, Fraser, Iain, and Haddock-Fraser, Janet. Using attribute importance rankings within discrete choice experiments: An application to valuing bread attributes. *Journal of Agricultural Economics*, 65(2):446–462, 2014a. ISSN 1477-9552. doi: 10.1111/1477-9552.12051. URL http://dx.doi.org/10.1111/1477-9552.12051.

Balcombe, Kelvin, Fraser, Iain, and McSorley, Eugene. Visual attention and attribute attendance in multi-attribute choice experiments. *Journal of Applied Econometrics*, 2014b. ISSN 1099-1255. doi: 10.1002/jae.2383. URL http://dx.doi.org/10.1002/jae.2383.

Bech, Mickael and Gyrd-Hansen, Dorte. Effects coding in discrete choice experiments. *Health Economics*, 14(10):1079–1083, 2005. ISSN 1099-1050. doi: 10.1002/hec.984. URL http://dx.doi.org/10.1002/hec.984.

Ben-Akiva, M.E. and Lerman, S.R. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press series in transportation studies. MIT Press, 1985. ISBN 9780262022170.

Campbell, Danny and Lorimer, Victoria. Accommodating attribute processing strategies in stated choice analysis: do respondents do what they say they do? In *European Association of Environmental and Resource Economists, 17th Annual Conference*, Amsterdam, 2009. URL http://www.webmeets.com/EAERE/2009/Prog/viewpaper.asp?pid=558.

Campbell, Danny, Hutchinson, W. George, and Scarpa, Riccardo. Incorporating discontinuous preferences into the analysis of discrete choice experiments. *Environmental and Resource Economics*, 41:401–417, 2008. ISSN 0924-6460. doi: 10.1007/s10640-008-9198-8. URL http://dx.doi.org/10.1007/s10640-008-9198-8.

Campbell, Danny, Hensher, David A., and Scarpa, Riccardo. Non-attendance to attributes in environmental choice analysis: a latent class specification. *Journal of Environmental Planning and Management*, 54(8):1061–1076, 2011. doi: 10.1080/09640568.2010.549367. URL http://dx.doi.org/10.1080/09640568.2010.549367.

Carlsson, Fredrik, Kataria, Mitesh, and Lampi, Elina. Dealing with ignored attributes in choice experiments on valuation of sweden's environmental quality objectives. *Environmental and Resource Economics*, 47:65–89, 2010. ISSN 0924-6460. URL http://dx.doi.org/10.1007/s10640-010-9365-6. 10.1007/s10640-010-9365-6.

Collins, Andrew T., Rose, John M., and Hensher, David A. Specification issues in a generalised random parameters attribute nonattendance model. *Transportation Research Part B: Methodological*, 56(0):234 – 253, 2013. ISSN 0191-2615. doi: http://dx.doi.org/10.1016/j.trb.2013.08.001. URL http://www.sciencedirect.com/science/article/pii/S0191261513001318.

Colombo, Sergio and Glenk, Klaus. Social preferences for agricultural policy instruments: joint consideration of non-attendance to attributes and to alternatives in modelling discrete choice data. *Journal of Environmental Planning and Management*, 57(2): 215–232, 2014. doi: 10.1080/09640568.2012.738190. URL http://dx.doi.org/10.1080/09640568.2012.738190.

Colombo, Sergio, Christie, Michael, and Hanley, Nick. What are the consequences of ignoring attributes in choice experiments? implications for ecosystem service valuation. *Ecological Economics*, 96(0):25 – 35, 2013. ISSN 0921-8009. doi: http://dx.doi.org/10.1016/j.ecolecon.2013.08.016. URL http://www.sciencedirect.com/science/article/pii/S0921800913002735.

Hensher, David. Attribute processing, heuristics, and preference construction in choice analysis. In *International Choice Modelling Conference*, Harrogate, UK, March 2009. URL http://www.icmconference.org.uk/index.php/icmc/icmc2009/paper/view/185.

Hensher, David A. How do respondents process stated choice experiments? attribute consideration under varying information load. *Journal of Applied Econometrics*, 21(6):861–878, 2006. ISSN 1099-1255. doi: 10.1002/jae.877. URL http://dx.doi.org/10.1002/jae.877.

Hensher, David A. Joint estimation of process and outcome in choice experiments and implications for willingness to pay. *Journal of Transport Economics and Policy*, 42(2):pp. 297–322, 2008. ISSN 00225258. URL http://www.jstor.org/stable/20054049.

Hensher, David A., Rose, John, and Greene, William H. The implications on willingness to pay of respondents ignoring specific attributes. *Transportation*, 32(3):203–222, 2005. ISSN 0049-4488. doi: 10.1007/s11116-004-7613-8. URL http://dx.doi.org/10.1007/s11116-004-7613-8.

Hensher, David A., Rose, John, and Bertoia, Tony. The implications on willingness to pay of a stochastic treatment of attribute processing in stated choice studies. *Transportation Research Part E: Logistics and Transportation Review*, 43(2):73 – 89, 2007. ISSN 1366-5545. doi: 10.1016/j.tre.2005.07.006. URL http://www.sciencedirect.com/science/article/pii/S136655450500058X.

Hensher, David A., Collins, Andrew T., and Greene, William H. Accounting for attribute non-attendance and common-metric aggregation in a probabilistic decision process mixed multinomial logit model: a warning on potential confounding. *Transportation*, 40(5):1003–1020, 2013. ISSN 0049-4488. doi: 10.1007/s11116-012-9447-0. URL http://dx.doi.org/10.1007/s11116-012-9447-0.

Hess, Stephane. Conditional parameter estimates from mixed logit models: distributional assumptions and a free software tool. *Journal of Choice Modelling*, 3(2):134–152, 2010. URL http://www.sciencedirect.com/science/article/pii/S1755534513700393.

Hess, Stephane. Impact of unimportant attributes in stated choice surveys. In *Transportation Research Board 91st Annual Meeting*, Washington, DC, 2012.

Hess, Stephane and Hensher, David A. Using conditioning on observed choices to retrieve individual-specific attribute processing strategies. *Transportation Research Part B: Methodological*, 44(6):781 – 790, 2010. ISSN 0191-2615. doi: http://dx.doi.org/10.1016/j.trb.2009.12.001. URL http://www.sciencedirect.com/science/article/pii/S0191261509001362. Methodological Advancements in Constructing Designs and Understanding Respondent Behaviour Related to Stated Preference Experiments.

Hess, Stephane and Hensher, DavidA. Making use of respondent reported processing information to understand attribute importance: a latent variable scaling approach. *Transportation*, 40(2):397–412, 2013. ISSN 0049-4488. doi: 10.1007/s11116-012-9420-y. URL http://dx.doi.org/10.1007/s11116-012-9420-y.

Hole, Arne Risa. A discrete choice model with endogenous attribute attendance. *Economics Letters*, 110(3):203 – 205, 2011. ISSN 0165-1765. doi: 10.1016/j.econlet.2010.11.033. URL http://www.sciencedirect.com/science/article/pii/S0165176510004064.

Hole, Arne Risa, Kolstad, Julie Riise, and Gyrd-Hansen, Dorte. Inferred vs. stated attribute non-attendance in choice experiments: A study of doctors' prescription behaviour. *Journal of Economic Behavior & Organization*, 96(0):21 – 31, 2013. ISSN 0167-2681. doi: http://dx.doi.org/10.1016/j.jebo.2013.09.009. URL http://www.sciencedirect.com/science/article/pii/S0167268113002412.

Holmes, Thomas P. and Adamowicz, Wiktor L. *A Primer on Nonmarket Valuation*, chapter Attribute-Based Methods, pages 171–219. Kluwer Academic Publishers, Dordrecht, 2003.

Hoyos, David. The state of the art of environmental valuation with discrete choice experiments. *Ecological Economics*, 69(8):1595 – 1603, 2010. ISSN 0921-8009. doi: http://dx.doi.org/10.1016/j.ecolecon.2010.04.011. URL http://www.sciencedirect.com/science/article/pii/S0921800910001540.

Kaye-Blake, William Henry, Abell, Walt L., and Zellman, Eva. Respondents' ignoring of attribute information in a choice modelling survey. *Australian Journal of Agricultural and Resource Economics*, 53(4):547–564, 2009. ISSN 1467-8489. doi: 10.1111/j.1467-8489.2009.00467.x. URL http://dx.doi.org/10.1111/j.1467-8489.2009.00467.x.

Kehlbacher, Ariane, Balcombe, Kelvin, and Bennett, Richard. Stated attribute non-attendance in successive choice experiments. *Journal of Agricultural Economics*, 64(3):693–706, 2013. ISSN 1477-9552. doi: 10.1111/1477-9552.12021. URL http://dx.doi.org/10.1111/1477-9552.12021.

Kosenius, Anna-Kaisa. Preference discontinuity in choice experiment: Determinants and implications. *The Journal of Socio-Economics*, 45(0):138 – 145, 2013. ISSN 1053-5357. doi:

http://dx.doi.org/10.1016/j.socec.2013.05.004. URL http://www.sciencedirect.com/science/article/pii/S1053535713000826.

Kragt, Marit E. Stated and inferred attribute attendance models: A comparison with environmental choice experiments. *Journal of Agricultural Economics*, 64(3):719–736, 2013. ISSN 1477-9552. doi: 10.1111/1477-9552.12032. URL http://dx.doi.org/10.1111/1477-9552.12032.

Lancaster, Kelvin J. A new approach to consumer theory. *Journal of Political Economy*, 74 (2):pp. 132–157, 1966. ISSN 00223808. URL http://www.jstor.org/stable/1828835.

Mariel, Petr, Marco, Boeri, and Hoyos, David. Dealing with controversial and non-attended attributes in discrete choice experiments. In *European Association of Environmental and Resource Economists, 19th Annual Conference*, Prague 2012. URL http://www.webmeets.com/EAERE/2012/prog/viewpaper.asp?pid=449.

Mariel, Petr, Hoyos, David, and Meyerhoff, Jürgen. Stated or inferred attribute non-attendance? a simulation approach. *Economía Agraria y Recursos Naturales*, 13(1):51–67, 2013. doi: 10.7201/earn.2013.01.03. URL http://purl.umn.edu/152821.

McFadden, Daniel. Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in econometrics*, chapter 4, pages 104–142. Academic Press, 1974.

Meyerhoff, Juergen and Liebe, Ulf. Discontinuous preferences in choice experiments: Evidence at the choice task level. In *European Association of Environmental and Resource Economists, 17th Annual Conference*, Amsterdam, June 2009. URL http://www.webmeets.com/EAERE/2009/Prog/viewpaper.asp?pid=545.

Payne, John W., Bettman, James R., and Johnson, Eric J. Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(3):534–552, 1988. ISSN 1939-1285(Electronic);0278-7393(Print). doi: 10.1037/0278-7393.14.3.534.

Puckett, Sean M. and Hensher, David A. The role of attribute processing strategies in estimating the preferences of road freight stakeholders. *Transportation Research Part E: Logistics and Transportation Review*, 44(3):379 – 395, 2008. ISSN 1366-5545. doi: http://dx.doi.org/10.1016/j.tre.2007.01.002. URL http://www.sciencedirect.com/science/article/pii/S1366554507000233.

Rose, John M., Hensher, David A., Greene, William H., and Washington, Simon P. Attribute exclusion strategies in airline choice: accounting for exogenous information on decision maker processing strategies in models of discrete choice. *Transportmetrica*, 8

(5):344–360, 2012. doi: 10.1080/18128602.2010.506897. URL http://dx.doi.org/10.1080/18128602.2010.506897.

Scarpa, Riccardo, Thiene, Mara, and Hensher, David A. Monitoring choice task attribute attendance in nonmarket valuation of multiple park management services: Does it matter?. *Land Economics*, 86(4):817 – 839, 2010. ISSN 00237639. URL http://le.uwpress.org/content/86/4/817.abstract.

Scarpa, Riccardo, Raffaelli, Roberta, Notaro, Sandra, and Louviere, Jordan J. Modelling the effects of stated attribute non-attendance on its inference: an application to visitors benefits from the alpine grazing commons. In *International Choice Modelling Conference*, 2011. URL http://www.icmconference.org.uk/index.php/icmc/ICMC2011/paper/view/381.

Scarpa, Riccardo, Zanoli, Raffaele, Bruschi, Viola, and Naspetti, Simona. Inferred and stated attribute non-attendance in food choice experiments. *American Journal of Agricultural Economics*, 95(1):165–180, 2013. doi: 10.1093/ajae/aas073. URL http://ajae.oxfordjournals.org/content/95/1/165.abstract.

Train, Kenneth. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2009.

Vuong, Quang H. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):pp. 307–333, 1989. ISSN 00129682. URL http://www.jstor.org/stable/1912557.

Weller, Priska, Oehlmann, Malte, Mariel, Petr, and Meyerhoff, JÃijrgen. Stated and inferred attribute non-attendance in a design of designs approach. *Journal of Choice Modelling*, 11(0):43 – 56, 2014. ISSN 1755-5345. doi: http://dx.doi.org/10.1016/j.jocm.2014.04.002. URL http://www.sciencedirect.com/science/article/pii/S175553451400027X. Process heuristics in choice analysis.

Chapter 4

A guide to heterogeneity features captured by parametric and nonparametric mixing distributions for the mixed logit model

Abstract

Unobserved heterogeneity is popularly modelled using the mixed logit model, so called because it is a mixture of standard conditional logit models. Although the mixed logit model can, in theory, approximate any random utility model with an appropriate mixing distribution, there is little guidance on how to select such a distribution. This study contributes to suggestions on distribution selection by describing the heterogeneity features which can be captured by established parametric mixing distributions and more recently introduced nonparametric mixing distributions, both of a discrete and continuous nature. We provide empirical illustrations of each feature in turn using simple mixing distributions which focus on the feature at hand.

4.1 Introduction

Heterogeneity in behaviour is widely acknowledged to be a fundamental aspect of discrete choice modelling (Hess et al., 2005; Desarbo et al., 1997; Allenby and Rossi, 1998), arising from differences in individual tastes, attitudes and perceptions, decision strategies, and other factors. Consequently, heterogeneity can affect many different parts of the choice model specification, including, for example, the taste parameters (taste variation), values of the attributes (perceptual variation), functional form (structural variation), and the error term (scale variation). The consequences of ignoring heterogeneity are similar to the consequences of other forms of misspecification: biased estimates and misleading policy implications (Desarbo et al., 1997).

Perhaps the most popular way of taking heterogeneity into account is the mixed logit model, which is defined by a choice probability expressed as a mixture of standard logit probabilities (Train, 2009):

$$p_{ij} = \int L_{ij}(\beta) f(\beta;\theta) d\beta \tag{4.1}$$

where p_{ij} is the probability that individual i will choose alternative j from a choice set, $f(\beta;\theta)$ is the density of the parameters β, θ are *hyperparameters* which parameterize the density of β[1], and L_{ij} is the standard choice probability for the conditional logit model

$$L_{ij}(\beta) = \frac{\exp(\beta' x_{ij})}{\sum_k \exp(\beta' x_{ik})}. \tag{4.2}$$

assuming a linear-in-parameters specification for the systematic utility. Inspection of Equation 4.1 reveals the origin of the name 'mixed logit': the mixed logit model is a mixture model where the components are conditional logit models and the mixing distribution is given by $f(\beta)$. Hence, the mixed logit model is also known as the 'mixed multinomial logit model' and the 'logit kernel model'.

The mixed logit model can be derived in a number of ways, but under the random utility framework, two interpretations are popular, the random parameters logit model and the error components logit model (Train, 2009). The former permits random coefficients in the systematic utility component of the random utility

$$U_{ij} = \beta_i' x_{ij} + \epsilon_{ij} \tag{4.3}$$

where x_{ij} is the vector of observed individual-specific and alternative-specific variables and β_i is the vector of individual-specific taste parameters, drawn from a density $f(\beta)$. The stochastic component solely consists of ϵ_{ij} under the random parameters formulation, and is iid Gumbel-distributed, leading to the logit choice probability expression Equation 4.1. The error components interpretation, on the other hand, fixes the taste parameters but permits additional error components creating arbitrary correlation between utilities for different alternatives

$$U_{ij} = \beta' x_{ij} + \mu_i' z_{ij} + \epsilon_{ij} \tag{4.4}$$

[1]For the rest of the text, we will use the term 'hyperparameter' to differentiate between parameters which enter the standard logit probability (parameter) and parameters which parameterize the mixing distribution (hyperparameter).

where μ is a vector of random coefficients and z_{ij} is a vector of observed variables, together comprising additional error components in random utility. Both formulations lead to the mixed logit model (Equation 4.1) and are theoretically equivalent, but they lend themselves to different interpretations. The random parameters formulation naturally lends itself to representing taste variation, since the random coefficients are interpreted as individual-specific taste parameters, while the error components formulation naturally lends itself to representing scale variation.

No matter the interpretation chosen, the mixed logit model has a special property: given the appropriate choice of variables and mixing distribution, it can approximate *any* random utility model to any degree of accuracy (McFadden and Train, 2000). This property is a double-edged blade: on the one hand, it makes the mixed logit model an extremely powerful model which can be used in almost any circumstances, but on the other hand, it implies that the mixed logit model can be incorrectly interpreted as revealing taste or scale variation when, in fact, it is simply capturing structures omitted through misspecification (Provencher and Bishop, 2004; Cherchi and Ortúzar, 2008; Cherchi and de Dios Ortúzar, 2010). For example, Cherchi and Ortúzar (2008) found in a simulation study that an alternative-specific random coefficient would give the impression of significant heterogeneity when in fact there was no taste variation in that term but rather correlation among alternatives. Consequently, analysts need to be careful drawing conclusions about taste variation when omitted structures may be present.

Furthermore, the result from McFadden and Train (2000) is only an existence result. It does not provide any guidance on which variables or which mixing distributions will achieve an arbitrarily accurate approximation (Train, 2008). In practice, most analysts use a continuous parametric distribution (Train, 2009), particularly the multivariate normal (Train, 2008). Although some studies test different continuous parametric distributions against each other (e.g. Marcucci and Gatta, 2012; Amador et al., 2008; Cherchi and Ortúzar, 2008), others simply adopt the normal distribution without further ado (e.g. Johnston and Duke, 2007; Birol et al., 2006; Horsky et al., 2006). However, the normal distributions has been strongly criticized for its strict assumptions, such as unbounded support, unimodality and symmetry. Occasionally, empirical applications of the mixed logit model use alternative mixing distributions which relax one or more of these assumptions, such as the log-normal distribution which is bounded on one side, and the latent class logit model which dispenses with parametric distributional assumptions altogether. Still, these alternatives have their own restrictions: the log-normal is afflicted with a heavy tail which may bias mean estimates, while the latent class logit model can usually only be estimated with a few classes. Not content with these simplistic alternatives, researchers have introduced a host of more advanced mixing distributions in the past decade. These include nonparametric continuous mixing distributions, which combine the benefits of avoiding parametric assumptions with a continuous representation of heterogeneity, and reparameterizations of the latent class logit model, which allow more detailed heterogeneity to be estimated with fewer parameters.

To the author's knowledge, there is no one study which comprehensively covers all the mixing distributions now available, old and new, parametric and nonparametric, continuous and discrete. Most of the relatively recent mixing distributions have seen limited use in empirical applications, possibly because many practitioners are not yet aware of them. On the other hand, those practitioners facing the full gamut of available mixing distributions may find specification an overwhelming task. In order to raise awareness of new mixing distributions and guide selection of mixing distributions, this study will describe mixing distributions in terms of the features of heterogeneity which they model. Although some studies have focused on one feature or another, this study is unique in its emphasis on features relevant to the research goal at hand and the coverage of multiple features. This study encourages practitioners to consider the *nature* of the heterogeneity which may be present, and choose the appropriate mixing distributions based on data, theory or policy relevance. For example, the data may reveal evidence of skew in the preference distribution, which can bias the mean from a policy standpoint. Another example comes from theory, which usually suggests that cost coefficients should be negatively signed, implying a distribution which is at least bounded on one side. Finally, if the policy question is about identifying different behavioural segments of customers (as in market segmentation, a fundamental concept in marketing), a preference distribution with multiple modes is indicated.

Wedel et al. (1999) points out that selecting an appropriate mixing distribution is especially difficult because specifying the form of parameter heterogeneity is largely an empirical issue. Yet his call for research into the 'theoretical underpinning of heterogeneity, with the purpose of identifying variables that need to be included in models and to assist researchers in the appropriate model specification' has not been answered in more than 10 years. However, he also suggests that 'empirical generalizations could help form theoretical foundations for the description of heterogeneity'.

Thus, the main contribution of the study is to comprehensively describe common and alternative mixing distributions, including raising awareness of the more recently introduced alternative mixing distributions, and with particular emphasis on the set of heterogeneity features which each mixing distribution captures. A secondary contribution of this study is an empirical search for each feature explored in this study using a case study in health economics, thus adding to the body of work which can form the empirical generalizations Wedel et al. (1999) call for.

First, we describe in section 4.2 the heterogeneity features which can help guide mixing distribution selection. We follow up in section 4.3 with a description of commonly used parametric mixing distributions with respect to the heterogeneity features they capture. In section 4.4, we illustrate an empirical search for those heterogeneity features using some of the discussed parametric mixing distributions. The combination of heterogeneity features, and in particular, more complex dependence structures, motivates the discussion of nonparametric mixing distributions in section 4.5, and another empirical illustration is provided in section 4.6. section 4.7 debates selecting between parametric and nonparametric mixing distributions, and section 4.8 concludes.

4.2 Heterogeneity features

Heterogeneity features are properties of the mixing distribution which may be relevant to theory or policy. We emphasize heterogeneity features as a guide for considering mixing distributions because they provide an approach for addressing unobserved heterogeneity from a basis of what is needed rather than what is found to fit best. Cherchi and de Dios Ortúzar (2010) suggest that statistical fit can be unhelpful as a mixing distribution selection criterion, because different specifications often fit equally well but result in very different predictions and policy implications. Furthermore, an emphasis on relevant heterogeneity features rather than statistical fit results in fewer mixing distributions for the analyst to test. A reliance on statistical fit as a selection criterion would imply that the analyst should estimate many (to the extent practical, all) available mixing distributions and select the one with best fit. In contrast, using heterogeneity features would involve identification of features deemed relevant under theory and policy and only estimating only the most parsimonious distribution which captures the relevant features. Indeed, if some of the relevant features are not observed empirically in the estimated mixing distribution, then an even more parsimonious mixing distribution may be chosen.

Below, we describe some heterogeneity features which can be used to select mixing distributions. The set of described heterogeneity features has largely been motivated by the reasons for dissatisfaction with the prevailing mixing distribution, the multivariate normal.

4.2.1 Range of support

The long tails of the normal distribution imply very positive and very negative values for taste parameters, which may be behaviourally implausible. For taste parameters which researchers have *a priori* sign expectations (*e.g.*, cost coefficient), support on both sides of zero will often imply implausibly large proportions of the distribution with the 'wrong' sign. Counterintuitive signs, particularly on cost coefficients, are a major driver of research into alternative mixing distributions (Hess et al., 2005; Train and Sonnier, 2005; Hess et al., 2006; Train, 2008; Rigby et al., 2009; Bastin et al., 2010; Campbell et al., 2010; Cirillo and Hetrakul, 2010; Hess, 2010; Chalak et al., 2012; Bastani and Weeks, 2013; Keane and Wasi, 2013). Finally, if the cost coefficient is specified as random with support which spans zero, WTP estimates will have infinite moments (Daly et al., 2012). Hensher and Greene (2003) and Hess et al. (2005, 2006) recommend bounded mixing distributions, emphasizing that the bounds should be estimated from the data. Once the mixing distribution is bounded, behaviourally implausible extreme values for taste parameters are eliminated, and coefficients with sign expectations are forced to be consistent with theory.

4.2.2 Number of modes

Economic theory rarely provides guidance on how many modes a mixing distribution should have. On the other hand, a central concept in marketing theory is market segmentation, the view that a heterogeneous market is composed of smaller homogeneous sub-markets (Wedel and Kamakura, 2000). The policy implication of multiple modes under a marketing context can easily be seen. Consider, for example, Campbell and Doherty (2013), which studied the demand for value-added services to chicken which improve food safety and quality (*e.g.*, food testing standards, traceability standards, animal health/welfare standards). They found that the demand for value-added services came from a niche market segment; while this segment was willing to pay a price premium for the services, other consumers were not willing to pay any price premium. A preference distribution failing to account for the multimodal nature of this preference distribution would have led to a misleading marginal WTP and revenue predictions. In particular, the niche market segment would have been charged a lesser price premium than they would actually be willing to pay, and the other consumers would not be willing to pay the price premium at all. Another way to think of the segmentation problem is that if one person wants a red balloon, and the other a blue balloon, a purple balloon satisfies no one. Under a multimodal distribution, the mode estimated under a unimodal distribution is unlikely to be located on any of the true modes.

However, accommodating multiple modes is not always necessary, even if they do exist. Suppose only population-level summary statistics, such as the population mean, are called for in a given application. Then a more restrictive mixing distribution with only one mode would still yield approximately the same measures as a more flexible mixing distribution which allows multiple modes.

4.2.3 Symmetry

Economic theory is also usually silent on whether a mixing distribution should be symmetric or asymmetric. Balcombe et al. (2011) motivates the consideration of skew with an example attribute, 'genetic modification', which could be associated with extreme disutility by some individuals but, at the same time, unlikely to be matched by extreme positive utility by other individuals. The authors then go on to develop a new mixing distribution which is a transformation of the normal distribution, but unlike previous transformations, can accommodate positive, negative or zero skew. In their empirical study of attitudes to bovine breeding technologies, they found significant evidence of skew, even after accounting for one potential source of skew, attribute non-attendance[2].

The presence of asymmetry, if unacknowledged, may bias the impression of where the 'middle' of the distribution is. Since the mean, as a measure of central tendency, is heavily influenced by extreme values, a heavy tail will pull the mean towards it, even if the bulk of the distribution is away from the mean. This point is illustrated in Figure 4.1, a schematic of symmetric and asymmetric distributions with the same mean. The shaded region represents the middle 50% of the asymmetric distribution, and the mean of both distributions is close to the edge of the shaded region.

Another policy implication of asymmetry is that there exists some individuals with relatively extreme preference parameter values, while the rest of the individuals have less extreme preferences. From a policy standpoint, it may be desirable to separate the two groups, so that the preferences of the 'mainstream' group can be served without the influence of the 'extreme' group, and the preferences of the 'extreme' group may or may not be served, or may be served with different policy instruments. This separation of preferences into 'majority/minority' is a special case of segmentation, mentioned in the discussion of number of modes, and, as the terminology would imply, could have important equity implications. There are different approaches to the issue of majority/minority preferences: identifying the majority/minority by modelling asymmetry allows for the full scope of these approaches, while ignoring asymmetry implicitly constrains the policy analysis to one particular approach.

[2]Attribute non-attendance is the behaviour of ignoring some attributes when performing a choice task. This behaviour is econometrically equivalent to a zero utility for the ignored attributes. If the true distribution of marginal utility for the attribute is symmetric, then the presence of attribute non-attendance would skew the distribution, since it confounds the true distribution with a point mass at zero.

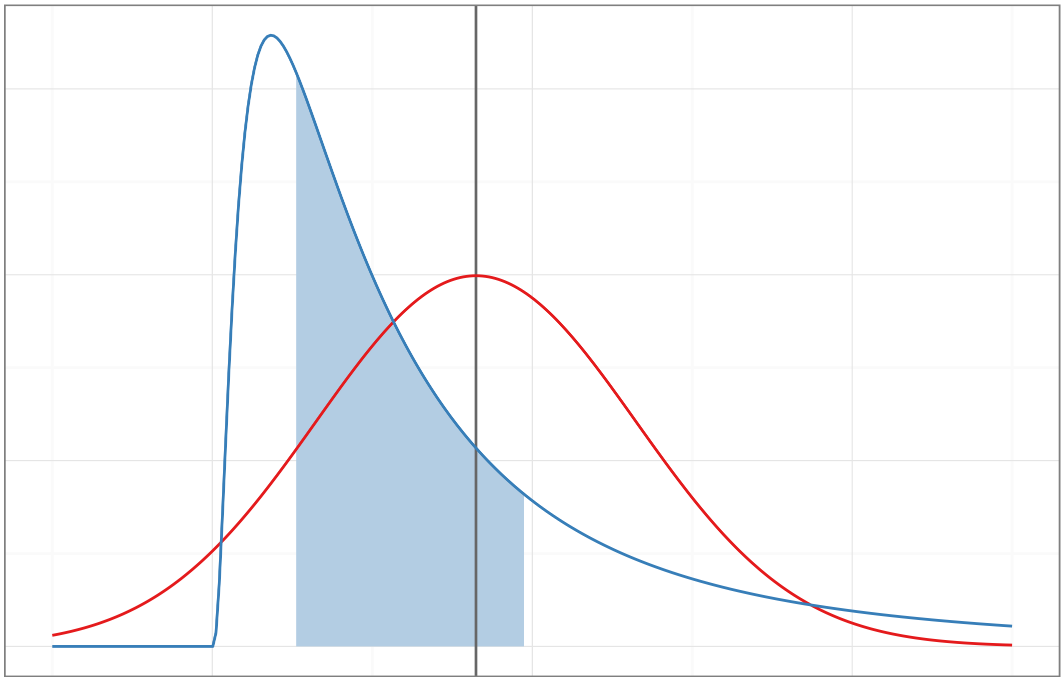


Figure 4.1: Schematic illustration of symmetric and asymmetric distributions with the same mean.

4.2.4 Dependence between random parameters

The conventional multivariate normal mixing distribution can theoretically accommodate a full correlation matrix between random parameters, but in practice, many applications assume mutual independence for tractability (Train and Sonnier, 2005; Hynes et al., 2008; Hess, 2014). Those applications which do permit full correlation have found the presence of significant correlation, better model fit, substantively different taste parameter estimates and important policy implications (Train and Sonnier, 2005; Hynes et al., 2008; Mabit et al., 2008; Rigby et al., 2009). Furthermore, correlation between random parameters is a consequence of scale heterogeneity (Hess and Rose, 2012). To see why, consider again the standard logit probability expression (Equation 4.2), except this time with the omitted scale parameter α present:

$$L_{ij}(\beta) = \frac{\exp(\alpha\beta' x_{ij})}{\sum_k \exp(\alpha\beta' x_{ik})}. \tag{4.5}$$

If scale heterogeneity is present, then α is individual-specific (*i.e.*, α_i). Since the scale parameter and the taste parameter are confounded and not separately identified (*i.e.*, only $\alpha\beta$ is identified), scale heterogeneity will scale all taste parameters simultaneously and equally across individuals. Thus, if taste heterogeneity is absent but scale heterogeneity is present, the random parameters will display perfect correlation. If scale heterogeneity is present but correlation is not permitted, the scale heterogeneity will manifest elsewhere, such as in the standard errors of the parameter estimates. Consequently, correlation between random parameters accommodates scale heterogeneity. As with taste heterogeneity, there is rarely an *a priori* reason not to suspect the presence of scale heterogeneity. However, correlation is only the simplest form of dependence between random parameters, and higher order relationships are also possible.

Both correlation and higher order relationships could have implications for policy design, because they suggest how attributes should be bundled to best effect. Dependence reveals opportunities for synergy between attributes (or warns of its opposite), and if the policy goal is to design programs or services based on multiple attributes, then the interactions between attributes can be exploited to design more effective programs. Even when the policy goal is to value one or two attributes, knowledge of interaction effects from other attributes can provide a more nuanced understanding of how attribute valuation may be contingent on other attributes.

4.3 Parametric mixing distributions

Alone, each of the heterogeneity features above can be captured by parametric mixing distributions. In this section we describe a selection of the most widely used parametric mixing distributions with respect to the heterogeneity features discussed in the previous section.

The multivariate **normal** distribution is probably the most widely used mixing distribution for MXLs. Its appeal can probably be attributed, in part, to the same reason which makes the normal distribution so common in more general usage: the central limit theorem. When the distributional form is unknown, researchers can at least appeal to the CLT to justify their usage of the normal distribution. Besides this general rationale, the normal distribution also provides advantages specific to the mixed logit model. Identifying the normal mixing distribution only requires two hyperparameters per random coefficient, and the interpretation of the hyperparameters is intuitive since they directly correspond to mean and standard deviation. Finally, estimating a mixed logit model with a normal mixing distribution is relatively fast and stable, and estimation routines are widely available in discrete choice modelling software packages. However, in terms of heterogeneity features, it is one of the most restrictive mixing distributions, since it assumes unbounded support, unimodality, symmetry and, frequently, mutual independence among random parameters.

The **log-normal** and **censored or truncated normal** are all transformations of the normal distribution which, most importantly, have a support bounded on one side. These distributions are often used by economists wishing to constrain the sign of the cost coefficient to respect economic theory. However, their heavy tails (on the unbounded side of their support) may bias mean estimates and their bounds are theoretically dictated rather than estimated from data. Furthermore, the log-normal distribution is notoriously difficult to estimate because of the resulting log-likelihood surface shape (Train and Sonnier, 2005). Like the normal, all three of these transformations are also unimodal, but unlike the normal, they are all asymmetrical. However, they are not any more flexible than the normal with respect to symmetry: they are necessarily asymmetrical, and do not permit symmetry. Hence, the presence of asymmetry is imposed rather than estimated when using these normal transformations.

In contrast, the **triangular** distribution permits negative, positive or zero skew. Therefore it permits both symmetry and asymmetry and is more flexible than the normal in this respect. In practice, however, the symmetrical variant of the triangular distribution is employed (Hess et al., 2006; Fosgerau and Hess, 2009; Hess, 2010), which is somewhat surprising since the triangular distribution may be the simplest distribution which offers flexibility with respect to symmetry. This flexibility comes at the price of one additional hyperparameter compared to the normal distribution, for a total of three. Like the normal and its transformations, the triangular distribution is unimodal. Unlike the normal and its transformations, support of the triangular distribution is bounded on both sides by bounds estimated from data.

Similarly, the **uniform** distribution is bounded on both sides by bounds estimated from data. Other than the one heterogeneity feature, the uniform distribution is otherwise extremely simple, for it does not have a mode (or, has infinitely many modes) and is perforce symmetrical.

In contrast, the **Johnson** S_B distribution is perhaps the most complicated of the commonly used parametric mixing distributions. It is, in fact, also a transformation of the normal distribution (Johnson, 1949), but involves two additional hyperparameters and can take on several shapes. Due to the greater number of hyperparameters, empirical identification can be problematic, and so many applications fix the additional hyperparameters. The additional hyperparameters are location parameters which dictate the bounds of the distribution; hence, the Johnson S_B is bounded with bounds that can either by fixed by the analyst or estimated from the data. The mean and standard deviation of the underlying normal distribution then become shape parameters which dictate whether the Johnson S_B is symmetrical or asymmetrical and unimodal or bimodal.

The parametric mixing distributions can all, in theory, be estimated with full correlation. When using a parametric mixing distribution, each random parameter is assigned its own distribution, which is tantamount to specifying the marginal distributions of the mixing distribution. Then, correlation between the marginal distributions can be induced using a decomposition of the variance-covariance matrix. However, correlation only captures linear relationships; if there are higher order relationships between the taste parameters, inducing correlation is not enough.

The alternative parametric distributions described above have been explored and compared against the multivariate normal in a number of studies (*e.g.* Hensher and Greene, 2003; Hess et al., 2005; Train and Sonnier, 2005; Hess et al., 2006; Fosgerau and Hess, 2009; Rigby et al., 2009; Cirillo and Hetrakul, 2010; Chalak et al., 2012), and in general, the bounded, more flexible distributions are preferred for their ability to avoid behaviourally implausible taste parameters and to capture complex features of the parameter distribution such as asymmetry. However, it is important to note that not all features may be relevant to the policy question at hand, nor present in the observed data.

4.4 Empirical illustration of parametric mixing distributions

In this section, we seek empirical evidence for each of the heterogeneity features described in section 4.2 using a dataset from a choice experiment on financial incentives for a behavioural weight loss program. Our baseline mixing distribution is the multivariate normal distribution with mutual independence, which could be considered the most frequently used, 'default' MXL mixing distribution. We then compare this baseline model against a set of MXL with alternative mixing distributions, chosen for their ability to simply capture each heterogeneity feature identified in section 4.2.

Previous empirical studies of mixing distributions have been dominated by the transportation, environmental and marketing domains, but this empirical illustration uses a choice experiment in health economics. The health domain is particularly interesting in this context because many of the heterogeneity features described previously are differently applicable to health preferences. For example, in transportation economics, signed cost coefficients have been a prime driver in considering bounded supports, but in health economics, cost/reward attributes may have counterintuitive signs because consumers may load them with other meanings beyond the purely financial. Thus bounded supports, or at least supports which have bounds not estimated from data, may be less desirable in a health context. Another example is multimodality: in the environmental context, benefit measures are important tools for policymaking, and in those cases, it is the benefit across large populations which are relevant. However, policymaking can often be more customized in the health context, due to the ability of individual programs, hospitals, and care providers to tailor their care to the particular subpopulation (which may not be representative of the general population) which utilizes their services, or which they hope to target with their services. Thus multimodality can be an actionable policy goal in the health context, allowing customization and targeting of desired subpopulations.

This study uses a choice experiment on the design of financial incentives for a weight loss program to illustrate common and alternative mixing distributions. The financial incentives were described using five attributes: reward amount, program location, payment form, reward condition and payment frequency. There were 1,296 respondents, each of which were presented with four choice tasks. A total of 96 choice sets were constructed according to the D-efficiency criterion.

4.4.1 Baseline distribution

The baseline model uses a normal mixing distribution with mutually independent random parameters. We specify the systematic utility to be a simple linear form, with attributes as the only terms. We specify the coefficients on reward amount and program locations to be random while the coefficients on the other attributes are left fixed. We limit the number of random parameters in order to facilitate clarity in comparison across the alternative distributions. We chose reward amount and program location to have random coefficients because previous work has indicated that are the two most important attributes, and focus group discussions have suggested heterogeneous preferences for their levels.

Table 4.1 displays the coefficient estimates for the baseline distribution. Note that all variables are effect-coded, except for reward amount, which is continuous. We can see that there is significant preference heterogeneity for reward amount and all program locations except for the community center.

4.4.2 Bounded alternative

Since we have a positive sign expectation on reward amount and no such expectations for the location levels, we expect boundedness to be a more acute need for the reward amount coefficient than the location coefficients. The bounded alternative distribution we use is the uniform mixing distribution. Like the normal mixing distribution, it is also described by two parameters, the minimum a and maximum b of the support. Alternatively, the uniform mixing distribution can also be parameterized by the center c

	Baseline model
Fixed parameters	
ASC	-0.16 (0.10)
Mag.0	0.09 (0.11)
Form: gym / cash	-0.14 (0.05)**
Form: medical / cash	-0.19 (0.05)***
Form: debit / cash	0.14 (0.05)**
Condition: weight / attendance	0.08 (0.05)
Condition: compliance / attendance	-0.05 (0.05)
Condition: attendance and compliance / attendance	-0.12 (0.05)*
Frequency: weekly / once	0.15 (0.05)**
Frequency: monthly / once	0.00 (0.05)
Frequency: quarterly / once	0.05 (0.05)
Random parameter means	
log(amount + 1)	0.35 (0.04)***
Location: workplace / clinic	-0.22 (0.05)***
Location: community center / clinic	0.18 (0.04)***
Location: church / clinic	-0.21 (0.05)***
Random parameter standard deviations	
s.log(amount + 1)	0.57 (0.03)***
s.Location: workplace / clinic	0.60 (0.10)***
s.Location: community center / clinic	0.24 (0.18)
s.Location: church / clinic	-0.65 (0.10)***
Log-likelihood	-4454.83
N	4994.00

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4.1: Coefficient estimates for baseline model, normal mixing distribution with mutually independent random parameters.

and spread s of the support. This alternative distribution is widely available in discrete choice modelling software.

	Baseline	Uniform
Fixed parameters		
ASC	-0.16 (0.10)	-0.19 (0.10)
Mag.0	0.09 (0.11)	0.06 (0.10)
Form: gym / cash	-0.14 (0.05)**	-0.14 (0.05)**
Form: medical / cash	-0.19 (0.05)***	-0.19 (0.05)***
Form: debit / cash	0.14 (0.05)**	0.14 (0.05)**
Condition: weight / attendance	0.08 (0.05)	0.08 (0.05)
Condition: compliance / attendance	-0.05 (0.05)	-0.05 (0.05)
Condition: attendance & compliance / attendance	-0.12 (0.05)*	-0.12 (0.05)*
Frequency: weekly / once	0.15 (0.05)**	0.15 (0.05)**
Frequency: monthly / once	0.00 (0.05)	-0.00 (0.05)
Frequency: quarterly / once	0.05 (0.05)	0.05 (0.05)
Random parameter means/centers		
log(amount + 1)	0.35 (0.04)***	0.34 (0.04)***
Location: workplace / clinic	-0.22 (0.05)***	-0.21 (0.05)***
Location: community center / clinic	0.18 (0.04)***	0.18 (0.04)***
Location: church / clinic	-0.21 (0.05)***	-0.22 (0.05)***
Random parameter standard deviations/spreads		
s.log(amount + 1)	0.57 (0.03)***	0.88 (0.04)***
s.Location: workplace / clinic	0.60 (0.10)***	-0.97 (0.17)***
s.Location: community center / clinic	0.24 (0.18)	0.48 (0.28)
s.Location: church / clinic	-0.65 (0.10)***	1.14 (0.15)***
Log-likelihood	-4454.83	-4469.75
N	4994.00	4994.00

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4.2: Coefficient estimates for baseline model and uniform mixing distribution.

Table 4.2 displays the coefficient estimates for both the baseline model and the uniform mixing distribution. We can see that the results are fairly similar across the fixed parameters and the random parameter means. However, the standard deviation and spread hyperparameters are not directly comparable. Instead, we visualize the two density functions (Figure 4.2).

We can see that across all random coefficients, the normal and uniform mixing distributions are quite similar to each other in terms of location and spread. Additionally, they

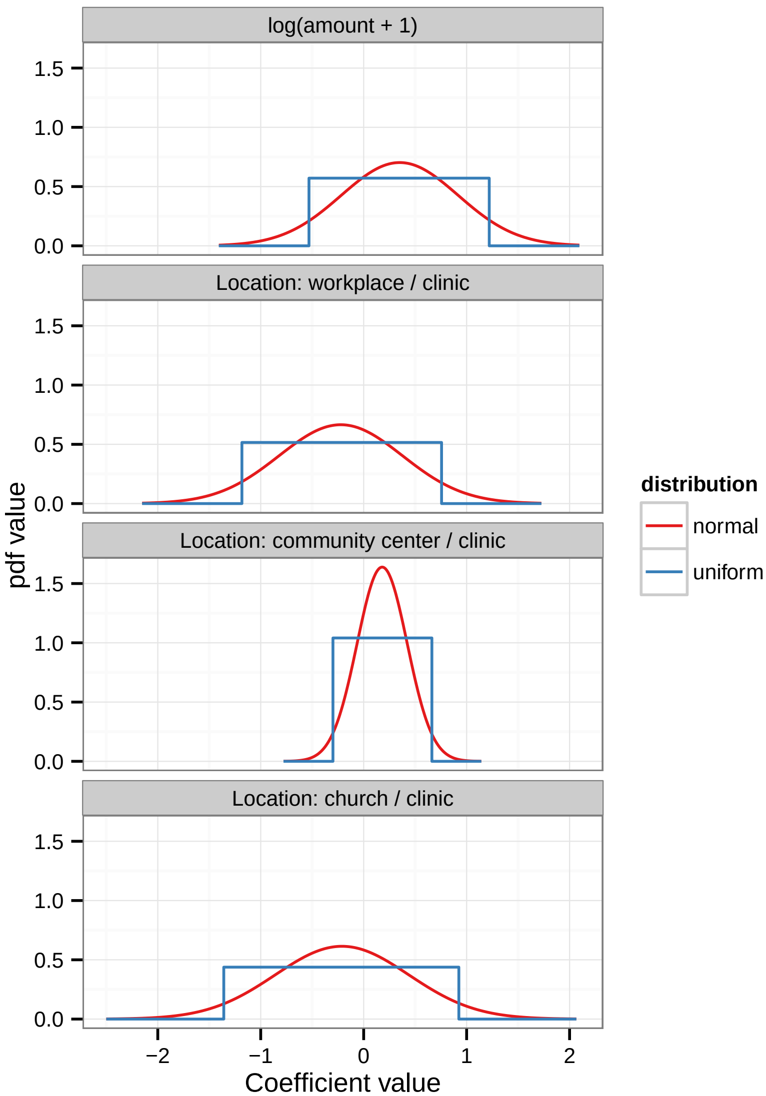


Figure 4.2: Density functions for estimated normal and uniform mixing distributions.

all display both negative and positive preferences. Some readers may find a noticeable proportion of reward amount preferences in the negative domain unusual. However, focus groups held during the development of the survey instrument indicated that some respondents felt offended at the presence of a financial incentive in a weight loss program. Moreover, financial incentives in weight loss programs have raised some degree of ethical controversy in both the popular press and academic literature, and there are several other examples of transactions for which monetization is culturally distasteful (Roth, 2007; Halpern et al., 2009; Schmidt, 2011; Lunze and Paasche-Orlow, 2013).

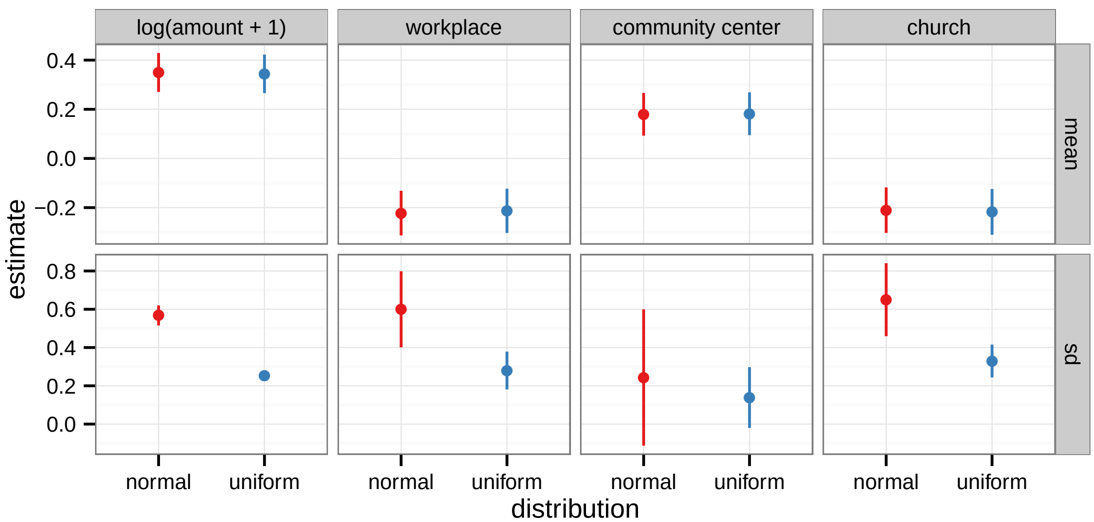


Figure 4.3: 95% confidence intervals for estimates of the mean and standard deviation of the normal and uniform mixing distributions.

We can also compare the estimated mixing distributions in terms of their central moments (Figure 4.3[3]). Given the obvious similarity in location of the densities in Figure 4.2, the overlapping confidence intervals for the mean estimates are unsurprising. The standard deviation estimates, on the other hand, are much more dissimilar, with the standard deviation of the uniform mixing distribution always lower than that of the normal mixing distribution. This bias is expected giving the bounded nature of the uniform distribution and the unbounded nature of the normal distribution.

Given the difference in standard deviation estimates between the normal and uniform

[3]Confidence intervals were computed using standard errors under the assumption of asymptotic normality. If the central moment was not a parameter directly estimated (*e.g.*, the standard deviation of the uniform mixing distribution), then the standard error for the central moment is calculated using the delta method.

Coefficient	Lower tail	Upper tail	Total
log(amount + 1)	0.06	0.06	0.12
Location: workplace / clinic	0.05	0.05	0.11
Location: community center / clinic	0.02	0.02	0.05
Location: church / clinic	0.04	0.04	0.08

Table 4.3: Amount of normal density outside of uniform density.

mixing distributions, we consider how much of the estimated normal density lies outside the estimated uniform density (Table 4.3). We find that proportions are quite modest, ranging from 5% to less than 15%. Since the means of the two distributions agree well with one another, this proportion is evenly split between the lower and upper tails.

Coefficient	Uniform	Normal	Difference
log(amount + 1)	0.30	0.27	0.03
Location: workplace / clinic	0.61	0.64	-0.03
Location: community center / clinic	0.31	0.23	0.08
Location: church / clinic	0.60	0.63	-0.03

Table 4.4: Amount of density in negative domain.

Finally, we consider the proportion of the estimated density in the negative domain, a useful statistic when the policy questions concerns determining how much of the population is for or against a particular attribute. This type of policy question is part of a broader category of questions which focus on a specific portion of the distribution, and may not be as forgiving to differences in the distribution as, say, the mean is. Table 4.4 reveals that again, the differences between the uniform and normal distributions are modest and not biased in a single direction. These results are in line with the suggestion in Sillano and de Ortúzar (2005) that using bounded distributions to enforce sign constraints are not altogether necessary. When considering the mean of the respondent-specific conditional coefficient distribution, they found that there were no respondents with the 'wrong sign' on the cost coefficient. Even under the unconditional distribution, the proportion of the distribution in the 'wrong domain' was quite small.

For this particular application, the estimated uniform and normal mixing distributions were very similar across most aspects we considered, although the standard deviation of the estimated uniform mixing distribution was smaller, a natural result of the uniform's bounded nature. These results suggest that unless policy questions are concerned with the extreme values of the distributions, policy interpretations in this case are likely to be quite similar whichever distribution is used.

4.4.3 Correlated alternative

Correlation between random parameters can provide important policy insights on which attributes and attribute levels work in synergy and which work in opposition to each other. Unanticipated correlation may produce unintended policy consequences, and so identifying the sign and significance of correlations between random parameters has policy relevance. The correlated alternative distribution we use is the correlated normal mixing distribution. In the baseline model, the random parameters are assumed to be mutually independent, but in the correlated normal mixing distribution, the random parameters are permitted to have an arbitrary variance-covariance matrix. In practice, the variance-covariance matrix is estimated by estimating the terms of the Cholesky factor in a Cholesky decomposition of the variance-covariance matrix (Croissant, 2013). Like the bounded alternative distribution, the correlated alternative distribution was also readily available in software packages.

Of most interest are the estimated covariance hyperparameters[4](Table 4.5). These indicate that there is no significant correlation in preferences between reward amount and location. However, there are significant negative correlations between location levels. In this case, preferences for different attributes are uncorrelated, but preferences for attributes within the same level are negatively correlated. In other words, an individual who prefers one location tends not to prefer the other locations. This result is intuitive, although we might expect some complementary location preferences if more location levels were available in the experiment. However, the location levels presented in the current application appear to be different enough to elicit preferences which view location levels more as substitutes.

Table 4.5 also indicates that the fixed parameter and random parameter means are quite similar to one another. The random parameter standard deviations cannot be directly compared in Table 4.5 because estimated standard deviations are displayed for the uncorrelated normal, but estimated variances are displayed for the correlated normal. Instead, we consider visualizations of both densities in Figure 4.4. We see that the estimated standard deviations of the uncorrelated distribution tend to be smaller than that of the correlated distribution, but only for the location level preferences. Taking into account the uncertainty of the estimated standard deviations (Figure 4.5) indicates that the difference is likely to be significant for the workplace and church but not the community center, owing to its large standard errors. Thus, the presence of significant correlations between random parameters would seem to negatively bias standard deviation estimates for the correlated random parameters.

[4]The covariance matrix hyperparameters were computed from the estimated Cholesky factor (as $\Sigma = C'C$), and the standard errors and Z-tests for significance of the hyperparameters were approximated using the delta method.

	Baseline	Correlated
Fixed parameters		
ASC	−0.16 (0.10)	−0.22 (0.11)
Mag.0	0.09 (0.11)	0.09 (0.11)
Form: gym / cash	−0.14 (0.05)**	−0.15 (0.05)**
Form: medical / cash	−0.19 (0.05)***	−0.20 (0.05)***
Form: debit / cash	0.14 (0.05)**	0.16 (0.05)**
Condition: weight / attendance	0.08 (0.05)	0.10 (0.05)
Condition: compliance / attendance	−0.05 (0.05)	−0.06 (0.05)
Condition: attendance & compliance / attendance	−0.12 (0.05)*	−0.12 (0.05)*
Frequency: weekly / once	0.15 (0.05)**	0.13 (0.06)*
Frequency: monthly / once	0.00 (0.05)	0.01 (0.05)
Frequency: quarterly / once	0.05 (0.05)	0.04 (0.05)
Random parameter means		
log(amount + 1)	0.35 (0.04)***	0.38 (0.04)***
Location: workplace / clinic	−0.22 (0.05)***	−0.25 (0.06)***
Location: community center / clinic	0.18 (0.04)***	0.18 (0.05)***
Location: church / clinic	−0.21 (0.05)***	−0.24 (0.05)***
Random parameter standard deviations		
s.log(amount + 1)	0.57 (0.03)***	
s.Location: workplace / clinic	0.60 (0.10)***	
s.Location: community center / clinic	0.24 (0.18)	
s.Location: church / clinic	−0.65 (0.10)***	
Random parameter covariance matrix terms		
amount.amount		0.36 (0.04)***
amount.workplace		0.01 (0.05)
amount.community		−0.02 (0.05)
amount.church		0.05 (0.05)
workplace.workplace		0.89 (0.25)***
workplace.community		−0.35 (0.12)**
workplace.church		−0.23 (0.12)*
community.community		0.49 (0.16)**
community.church		−0.25 (0.13)
church.church		0.88 (0.21)***
Log-likelihood	−4454.83	−4435.63
N	4994.00	4994.00

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4.5: Coefficient estimates for baseline model and correlated normal mixing distribution.

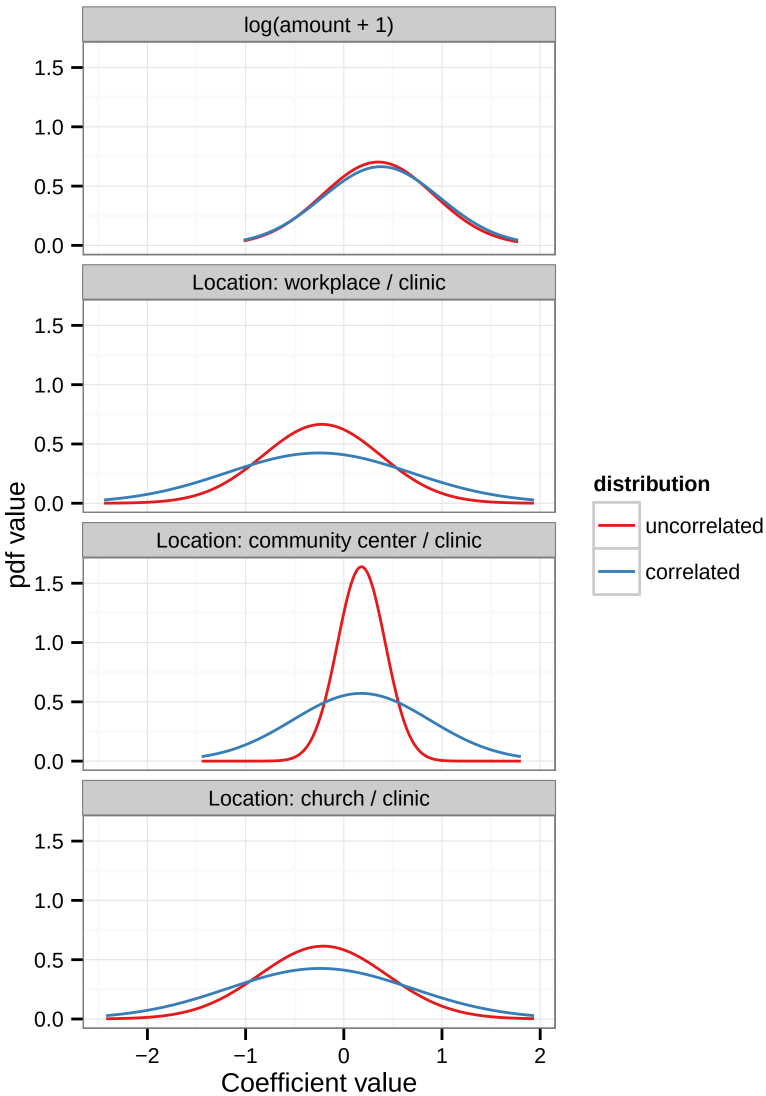


Figure 4.4: Density functions for estimated uncorrelated and correlated mixing distributions.

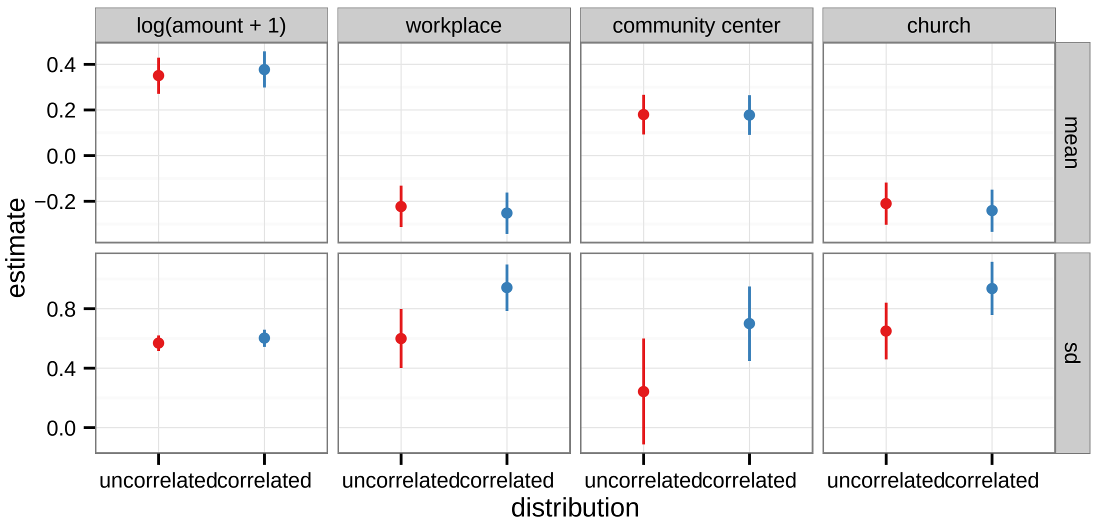


Figure 4.5: 95% confidence intervals for estimates of the mean and standard deviation of the uncorrelated and correlated mixing distributions.

Using a mixing distribution which permits correlation, we uncovered significant correlation between random coefficients for levels with the same attribute, as well as evidence of bias in the estimated standard deviation for those random coefficients affected by correlation. However, correlation only captures linear dependence. Nonlinear dependence may be present, but in practice, dependent random variables are generated using the Cholesky decomposition of a covariance matrix. Possible approaches to assessing the presence of higher order relations between random coefficients is using copulas to describe the dependence relation between parametric random coefficients, using a mixture of distributions, or using a nonparametric discrete mixture distribution which directly estimates the joint distribution and hence the dependence structure.

4.4.4 Asymmetrical alternative

The presence of asymmetry may have important policy implications for using mean estimates, since a long tail can skew the mean of distribution. The policy relevance of asymmetry is related to the policy relevance of the discrete, nonparametric alternative mixing distribution: in the latter case, we tried to reveal minority preferences. In the former case, we are trying to prevent minority preferences from skewing our understanding of the majority preferences. The asymmetrical alternative distribution we use is the triangular distribution, parameterized by three hyperparameters, the lower bound a, the upper bound b and the mode c (or equivalently, the left-hand spread s_1, the right-hand spread s_2 and the mode c). Most empirical applications of the triangular distribution use the symmetrical variant (in which c is fixed to $c = \frac{a+b}{2}$), but in its most general form, the triangular distribution may have positive, zero or negative skew.

In addition to most empirical applications, most estimation routines of the triangular distribution also implement only the symmetrical variant. Consequently, we again coded our own simulated likelihood function (Train, 2009) and maximized it using the maxLik package (Henningsen and Toomet, 2011). We again found it necessary to multistart the routine to avoid being trapped in local maxima. We also implemented three methods for simulating draws from the triangular distribution: the inverse cdf method, the method proposed by Dekker (2014), and the method proposed by Stein and Keblis (2009).

Unfortunately, we found empirically identifying all random parameters difficult. Despite using multiple starting values and methods for drawing from the triangular distribution, we were unable to estimate a model in which more than one of the random parameters was identified. The model with the highest likelihood value was one where the random coefficient on reward amount was identified.

In Figure 4.6, we present the density functions of the estimated mixing distributions (see section 6.H for table of coefficients). The lack of identification in the location parameters is visually apparent. The estimated mixing distributions for the reward amount coefficient, however, overlap to a large extent. The estimated triangular distribution displays a modest negative skew, which is in contrast to the LCL results. Under the LCL specification, the reward amount pmf displayed a positive skew, with class 3 having a strong positive preference for increasing reward amount unmatched by a strong negative preference from any of the other classes. Speculating which direction the skew should take is difficult, since either argument is believable: 1) that individuals offended by a financial reward for weight loss have strong negative preferences for increasing reward amount, unmatched by equally strong positive preferences from those who are not offended, or 2) that the motivation some individuals feel for monetary rewards is stronger than the distaste others feel for those some rewards.

However, when we take into account the standard error on the skewness of the reward amount coefficient, we see that the negative skew is not significant (Figure 4.7). Considering the other central moments, we find that the mean of the triangular is similar to the mean of the normal across all random coefficients. Thus, the identification problem seems to affect the spread of the triangular distribution rather than the location. Accordingly, the estimated triangular standard deviations are lower than the estimated standard deviation for the normal mixing distribution across all random coefficients, including the reward amount coefficient. This last observation is similar to the result we found when comparing the uniform and normal mixing distributions; due to the bounded nature

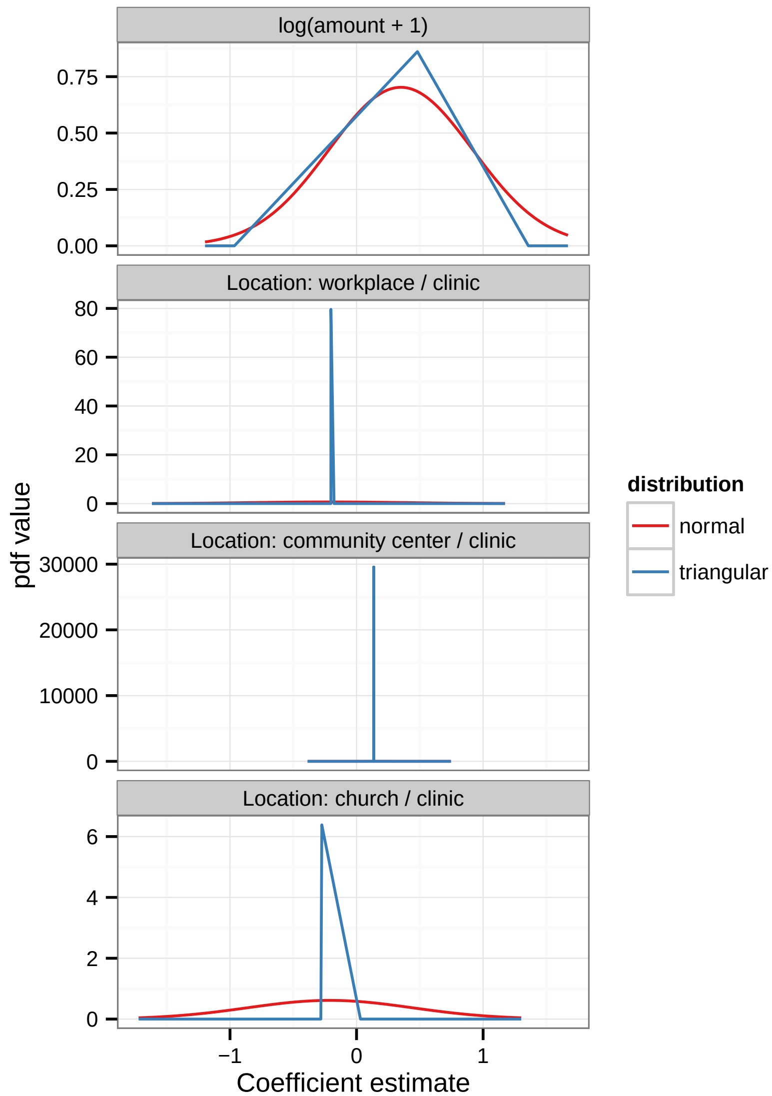


Figure 4.6: Density functions of normal and triangular mixing distributions.

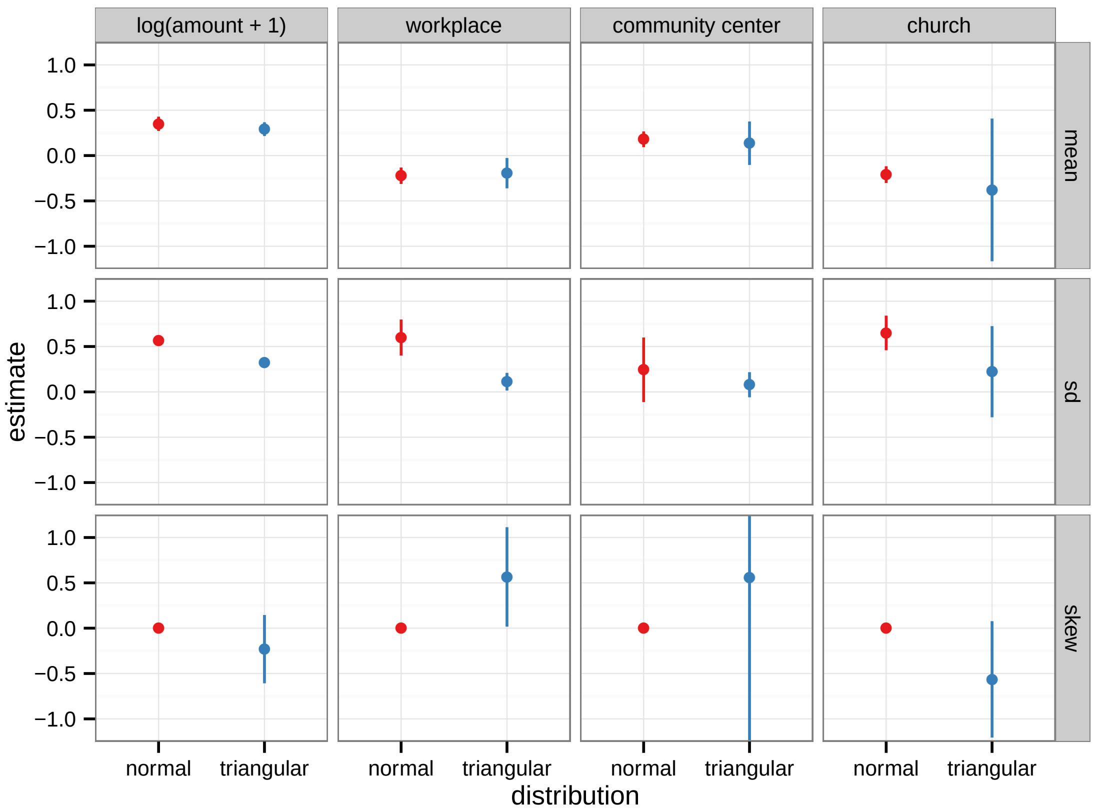


Figure 4.7: 95% confidence intervals for estimates of the mean, standard deviation and skew of the normal and triangular mixing distributions.

of the triangular and uniform mixing distributions, they have lower estimated standard deviations even when they appear to overlap the normal mixing distribution to a large extent. The lack of identification in spread also affects the estimates of skew in the triangular distribution; the standard errors or the skew estimates are very large and make inference of the presence and directionality of skew difficult.

4.4.5 Bimodal alternative

When motivating the discrete, nonparametric mixing distribution, we argued for the importance of identifying minority preferences. Although the latent class logit model is one method for doing so, identifying mode location(s) is another approach which can also do so. In contrast to the latent class logit model, multimodal mixing distributions are more likely to identify the highest modes, representing [possibly] multiple 'mainstream' preferences rather than revealing 'extreme' preferences. The multimodal alternative distribution we use is the Johnson S_B distribution, which can accommodate either one or two modes. The Johnson S_B is defined as a transformation of the standard normal distribution (Johnson, 1949),

$$x = \xi + \lambda \frac{\exp(\frac{z-\gamma}{\delta})}{1 + \exp(\frac{z-\gamma}{\delta})} \tag{4.6}$$

where x is the Johnson S_B random variable, z is a standard normal random variable, γ and δ are shape parameters representing the mean and standard deviation of the underlying normal random variable, and ξ and λ are location and scale parameters, respectively describing the minimum and the spread of the distribution support. The relative values of the shape parameters γ and δ determine whether the distribution is unimodal or bimodal[5].

Although the Johnson S_B is a relatively common alternative mixing distribution for the mixed logit model, its availability in software packages was limited. As with the LCL and the general triangular distribution, we coded our own likelihood function and used the R package maxLik (Henningsen and Toomet, 2011) to maximize it. As usual, we multistarted the estimation routine[6] and simulated the Johnson S_B distribution as a transformation of the standard normal distribution.

As with the general triangular distribution, we found empirical identification to be an practical obstacle. Applications of the Johnson S_B distribution to the mixed logit model frequently fix the location and scale parameters ξ and λ in order to improve identification. However, in this case, identification remained elusive even after reducing the number of free parameters.

Moreover, identification of bimodality under fixed location and scale may be especially difficult because of the inflexibility of the bimodal Johnson S_B shape. When bimodal, the modes occur only at the boundaries of the distribution support. Under fixed location and scale, the analyst would have to guess the locations of the modes in order to ascertain their existence with model estimates. Thus, bimodality is ideally established by simultaneously estimating all four hyperparameters.

Figure 4.8 presents the estimated density functions for the Johnson S_B distribution, with all four hyperparameters freely estimated (see section 6.I for table of coefficients). As with the triangular distribution, the reward amount coefficient appears to be identified,

[5]Johnson and Kitchen (1971) give the specific condition for bimodality as $|\gamma| < \frac{1}{\delta}\sqrt{1-2\delta^2} - 2\delta \tanh^{-1}(\sqrt{1-2\delta^2})$.

[6]Starting values for the location and scale parameters ξ and λ where taken from the bounds of a uniform mixed logit, estimated on random partitions of the data. Starting values for the shape parameters γ and δ were drawn from a uniform distribution on $[-2, 2]$ and $[0, 2]$, respectively.

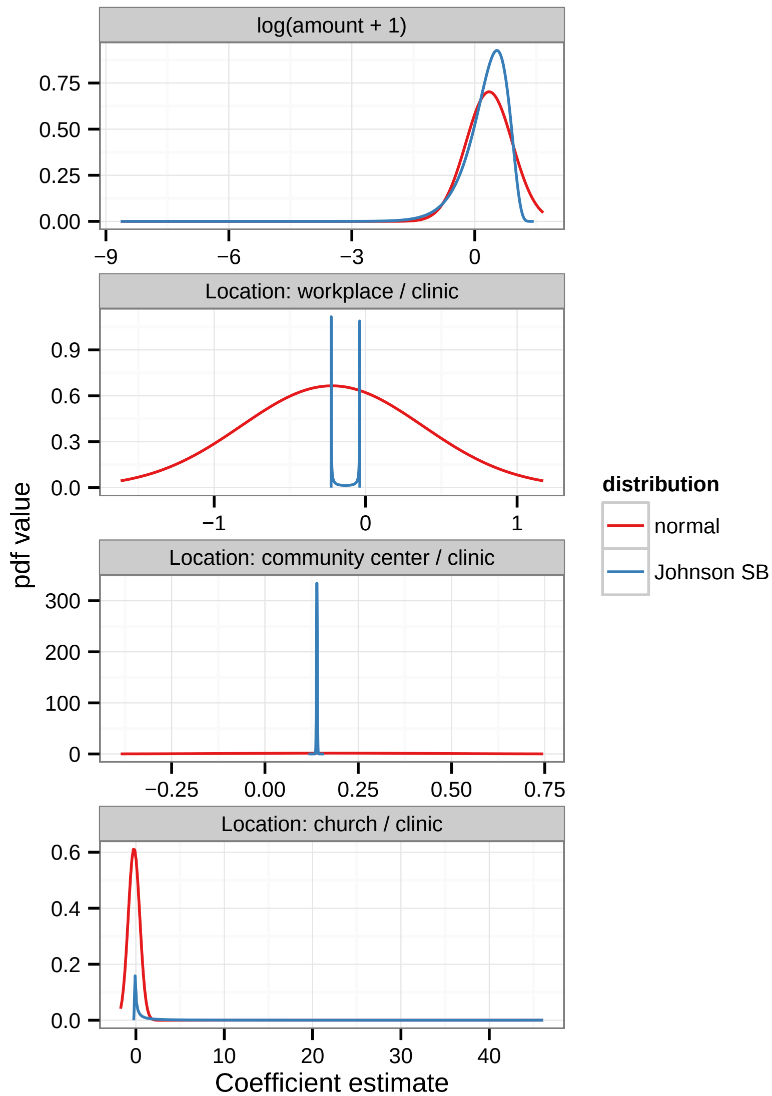


Figure 4.8: Density functions for estimated Johnson S_B and normal mixing distributions.

with the Johnson S_B density largely overlapping the normal density. Interestingly, the Johnson S_B distribution displays more pronounced negative skew than the triangular distribution. The workplace coefficient is the only random parameter which suggests evidence of bimodality.

We do not formally compare the moments of the estimated Johnson S_B distributions to the normal distribution. The moments are fairly complex to compute, and their partial derivatives (for use in the delta method) even more so. Furthermore, Hess et al. (2005) have pointed out that the performance of the Johnson S_B estimating moments correctly is highly dependent on the underlying true distribution; hence the moments may not be reliable and a comparison with the normal moments may not be meaningful.

4.5 Nonparametric distributions

In section 4.4, we sought empirical evidence for each of the heterogeneity features identified in section 4.2. We found only weak evidence for multimodality and number of modes, and for the most part, policy implications remained similar under mixing distributions with unbounded and bounded support. However, strong evidence for correlation between random parameters was found. Since correlation is only the simplest form of dependence between random parameters, the question remains whether there exists a more complex dependence structure.

Since the current state of the art cannot capture more complex dependence structures using parametric mixing distributions, we turn to nonparametric mixing distributions[7], in particular the discrete nonparametric mixing distributions. In this section, we first discuss the discrete nonparametric mixing distributions, then the continuous nonparametric mixing distributions. Although the ability of the latter to capture dependence between random parameters is limited, they are still useful for capturing combinations of heterogeneity features not available under parametric mixing distributions. Although our empirical illustration did not reveal the need for, *e.g.*, capturing multimodality and asymmetry simultaneously, such a situation may be possible in another case. Continuous parametric mixing distributions can answer such a need.

[7]The term 'nonparametric' is used loosely here and elsewhere in the paper to mean any distribution which is not strictly parametric, thus encompassing semiparametric and seminonparametric approaches as well as fully nonparametric techniques.

4.5.1 Discrete nonparametric distributions

The latent class logit model is a well-established example of a nonparametric mixing distribution which has found favour in multiple disciplines because it is intuitive and easy to interpret. The latent class logit model is an example of a finite mixture model, in which the mixing distribution is discrete and there are a finite number of support points. Thus, the number of support points limits the richness of preference heterogeneity which can be described by the latent class logit model. The more support points in the distribution, the higher the resolution, so to speak, of the distribution, thus more clearly revealing potential heterogeneity features such as symmetry and modality. In practice, the number of support points which can be estimated is relatively small since the number of hyperparameters grows linearly with the number of support points specified, and estimation issues are usually encountered when there are too many hyperparameters. As a result, variants of the finite mixture theme have been introduced in order to increase the number of support points which can be estimated. We first describe the original latent class logit model and then describe the variants which have recently been introduced.

Latent class logit model

The **latent class logit model** can be described behaviourally as follows: suppose there are S latent segments in the population. Within the segments, preferences are homogeneous, but across segments, preferences may be heterogeneous. Following the random utility framework, the probability that individual i will choose alternative j, conditional on being a member of segment s, is

$$p_{ij|s} = \frac{\exp(\beta_s' x_{ij})}{\sum_k \exp(\beta_s' x_{ik})} \tag{4.7}$$

The segments are unobservable (hence latent), and so an individual's segment membership is only known up to a probability. The unconditional probability that individual i will choose alternative j is

$$p_{ij} = \sum_s \pi_s \frac{\exp(\beta_s' x_{ij})}{\sum_k \exp(\beta_s' x_{ik})} \tag{4.8}$$

where π_s is the probability that the individual belongs to segment s.

Econometrically, the latent class logit model can be described as a mixed logit model with a finite mixing distribution. The class membership probabilities and class-specific preference parameters are the probability masses and mass point locations, respectively, of a probability mass function (pmf) describing the distribution of a discrete random variable. This perspective is apparent when comparing Equation 4.1 to Equation 4.8; the former is a continuous mixture with an integral and the latter is a finite mixture with a sum. This contrast is the reason the latent class logit model is seen as the discrete analog of the typical random parameters logit model, which is specified with a continuous, parametric distribution.

The flexibility of the latent class logit model grows with the number of classes. Since the number of hyperparameters grows linearly with the number of classes, classes are typically limited to less than a dozen. If the analyst specifies too many classes, then the estimation will frequently fail to converge or experience other problems. Consequently, variants on the latent class logit model which are more parsimoniously parameterized, such as the mass point MXL and the fixed point MXL, have been introduced.

Mass point mixed logit model

The **mass point MXL** is essentially a reparameterization of the latent class logit model. It was first introduced in Dong and Koppelman (2003) and later compared against other mixing distributions in simulation studies and empirical applications (Hess et al., 2007; Campbell et al., 2010; Dong and Koppelman, 2014). In this model, the number of support points is parameter-specific. Let β denote the vector of random parameters and p the index of the random parameter within the vector. Then the pth random parameter, β_p, has a parameter-specific number of marginal support points, which we denote by M_p. Each marginal support point is associated with a probability mass π_{mp}, where m indexes the support point. In order to be proper probabilities, π_{mp} must satisfy $0 \leq \pi_{mp} \leq 1$ and $\sum_{m=1}^{M_p} \pi_{mp} = 1$. Thus, the total number of possible joint support points is $M \equiv \prod_p M_p$.

To ease notation, let $m = 1, \ldots, M$ index the joint support points. Let π_m be the probability mass associated with each joint support point, acknowledging that π_m is derived as a product of the relevant marginal probabilities π_{mp} across parameters p. Under this model, the probability that individual i will choose alternative j is

$$p_{ij} = \sum_m \pi_m \frac{\exp(\beta_m' x_{ij})}{\sum_k \exp(\beta_m' x_{ik})} \tag{4.9}$$

where β_m is a vector corresponding to the m^{th} joint mass point.

The distinction between the mass point MXL and the latent class logit model is subtle, and lies entirely in how many support points can be specified with a given number of hyperparameters. In the latent class logit model, the analyst specifies the number of segments S, and each segment corresponds to a *joint* support point. Thus, there are as many joint support points as there are segments. In the mixed point MXL, the analyst specifies the number of *marginal* support points for each parameter M_p, and so the number of segments is equal to the product of the number of support points for each parameter, $\prod_p M_p$. Consequently, the same number of joint support points can be specified using a different number of hyperparameters. For the latent class logit model, each segment requires P location estimates and one probability mass estimate, so $S \times P + S$ estimates are needed for S joint support points. For the mass point MXL, only the location and probability mass of marginal support points need to be estimated, so $2\sum_p M_p$ estimates yield $\prod_p M_p$ joint support points. Visually, joint support points in the mass point MXL are defined on a P-dimensional lattice, whereas joint support points in the latent class logit model can be defined anywhere in P-dimensional space; the freedom gained comes at the cost of additional hyperparameters.

Fixed point mixed logit model

In the latent class logit model and mass point MXL, both the locations of and the probability masses at each support point are freely estimated. In this next model, the locations of the support points are fixed, and only the probability mass at each point is estimated. Accordingly, we term this model the **fixed point MXL**. The choice probability expression is identical to Equation 4.8 and Equation 4.9, but the location of the support points β_m are now specified by the analyst rather than estimated.

The fixed point MXL was first introduced by Bajari et al. (2007), and compared against other mixing distributions in simulation studies and empirical applications by Train (2008) and Bastani and Weeks (2013). The chief benefit of the fixed point MXL compared to the other discrete mixing distributions is the ability to estimate many more joint mass points. Whereas latent class logit models and mass point MXL often fail to converge or yield degenerate solutions[8] when the number of support points is only a handful, the fixed point MXL is fast and stable even when estimating probability masses for hundreds of thousands support points (Train, 2008). However, the results are sensitive to the analyst's specification of the support point locations, and in particular the range of the parameter space. In response to this weakness, Bastani and Weeks (2013) developed some heuristics to address range selection.

Fixed mass mixed logit model

A natural corollary to the fixed point MXL is the **fixed mass MXL**, in which the probabilities masses are fixed but the support points are estimated. Although this model has not been introduced explicitly in the literature as an approach for estimating MXL mixing

[8]Solutions in which probability masses are close to zero for some segments or some segment locations are very close together.

distributions, Campbell et al. (2009) exploited the idea to detect 'outlier' preferences. In their application, the probability masses were fixed at low values (*e.g.*, 1%) in order to identify the extreme ends of the distribution. This approach could clearly be extended to estimate locations of many probabilities masses rather than just the extremes.

All these variants on the latent class logit model have similar properties. They all have bounded support, since the number of support points is specified by the analyst and therefore finite. They can accommodate multiple modes, asymmetry and arbitrary dependence between random parameters because the joint distribution is directly estimated. We will expand on this last point because it is less obvious than the other properties.

Each mixing distribution described above directly estimates the joint mixing distribution. As we have previously mentioned, estimating a parametric mixing distribution involves estimating the marginal distribution of each dimension (*e.g.*the distribution of an individual random coefficient) and then imposing a correlation structure among these marginal distributions. In contrast, the discrete nonparametric mixing distributions directly estimate the joint mixing distribution, and the dependence structure is implicitly estimated because it is a byproduct of the joint distribution (Provencher et al., 2002; Provencher and Moore, 2006; Hess et al., 2011; Hess, 2014). Consequently, parametric mixing distributions are limited to analyst-specified dependence structures (in practice, correlation only), while the discrete nonparametric mixing distributions can reveal arbitrarily complex higher order relationships.

4.5.2 Continuous nonparametric distributions

The mass point MXL, fixed point MXL and fixed mass MXL arose in response to the limited heterogeneity which traditional latent class logit models could capture. At the same time, a separate branch of the literature reacted to the same problem in a different way, by using *continuous* nonparametric mixing distributions rather than discrete ones.

Mixture of distributions mixed logit model

The most obvious approach for increasing the heterogeneity captured by the latent class logit model is to extend it so that each segment has random rather than fixed taste parameters. This model has been called by different names, including random parameters latent class logit model and latent class mixed multinomial logit model, which recognize its origins with the latent class logit model (Bujosa et al., 2010; Campbell and Doherty, 2013; Greene and Hensher, 2013). However, the formulation is equivalent to specifying the mixing distribution as a **mixture of distributions** itself (Dong and Koppelman, 2003;

Train, 2008; Fosgerau and Hess, 2009; Campbell et al., 2010, 2014; Fosgerau, 2014). That is,

$$F(\beta) = \sum_s \pi_s G(\beta; \theta_s) \tag{4.10}$$

where $F(\beta)$ is the cdf of the MXL mixing distribution, and $G(\beta; \theta_s)$ are components of a finite mixture with segment-specific hyperparameters. Setting the base distribution to be normal, $G(\cdot) = \Phi(\cdot)$, is particularly appealing because any continuous distribution can be approximated arbitrarily well by a finite mixture of normals. A further advantage of using a mixture of distributions is the ability to accommodate point masses, since the distribution within any given segment may become degenerate. Thus, the mixture of distributions can represent both discrete and continuous types of heterogeneity. The complexity of this model increases with the number of mixtures included, and most applications use only two or three components.

Because the mixture of distributions MXL is a generalization of the latent class logit model, it shares many of the same properties. For example, it can accommodate multiple modes, with the number depending on the base distribution and number of mixture components. It can also accommodate complex dependence structures, not only in the same way that the latent class logit model can, but also by permitting full correlation between random coefficients within each mixture component. Although doing so would appear to rapidly increase the number of parameters to be estimated, perhaps beyond what could be supported by many datasets, Train (2008) was successful in doing so in his empirical application of the mixture of distributions model. The one property which is different between the latent class logit model and the mixture of distributions MXL is boundedness. Unlike the latent class logit model, the mixture of distributions MXL is not necessarily bounded. The support of the mixture of distributions MXL depends on what distributions are used as components of the mixture; if the normal is used, as is always the case in practice, then the support will be unbounded.

In fact, the mixture of distributions model is an example of a sieve estimator, defined as an estimator which approximates unknown functions with a series of basis functions. In the case of the mixture of distributions MXL, the basis functions are the base distribution of the mixture, such as the normal in a mixture of normals. The quality of the approximation depends on the basis functions and the number of terms in the series. If the basis function is a good approximation to the unknown function, then only a small number of terms should be necessary (Fosgerau, 2014). The method of sieves has been a popular approach for developing nonparametric MXL mixing distributions, and below we describe other sieve estimators which have also been proposed.

Sieve estimator with Legendre polynomials

Fosgerau and Bierlaire (2007) introduced a **sieve estimator with Legendre polynomials** as the basis function. The series of Legendre polynomials seminonparametrically approximate the derivative of a transformation function, rather than the mixing distribution directly. To be more precise, let $F(x)$ be the true cdf of the random coefficient and $G(x)$ be a base distribution which is selected *a priori* . Then the true cdf can be expressed as a transformation of the base distribution:

$$F(x) = Q(G(x)) \tag{4.11}$$

Since $Q : [0,1] \to [0,1]$ and monotonically increases, it satisfies the conditions to be a cdf. Let q be the corresponding probability density function (pdf), which can be seminonparametrically approximated as a series of Legendre polynomials

$$q(x) \approx \frac{1}{K}\left(1 + \sum_{k=1}^{N} \delta_k L_k(x)\right)^2 \tag{4.12}$$

where K is a normalizing constant to ensure that $q(x)$ integrates to 1, N is the number of Legendre polynomials used in the finite approximation, δ_k are unknown coefficients to be estimated and L_k is the k^{th} Legendre polynomial. The true distribution can then be recovered as $f(x) = q(G(x))g(x)$.

The base distribution $G(x)$ serves two purposes: determining the support of the true density and providing an initial guess for the true distribution $F(x)$ (Bierens, 2008). If the guess is right, $F(x) = G(x)$, then the transformation $Q(\cdot)$ will be the cdf for the uniform distribution. As with other sieve estimators, the more terms there are in the series expansion, the more flexible the nonparametric mixing distribution is. In practice, three or fewer terms are usually sufficient (Scarpa et al., 2008; Fosgerau and Hess, 2009). Fosgerau and Bierlaire (2007) originally developed this approach as a test for the choice of mixing distributions: the significance of the higher order terms indicates whether the parametric base distribution is appropriately specified or not.

Sieve estimator with B-splines

Bastin et al. (2010) introduced a different **sieve estimator with B-splines** as the basis functions. In this approach, the spline function approximates the inverse cdf of a random coefficient, hence the domain is $[0,1]$. The knot locations are fixed on $[0,1]$ by the analyst and the coefficients on the basis functions are estimated. The spline function is guaranteed to be nondecreasing (and hence a proper inverse cdf) if the coefficients on the basis functions are also nondecreasing, a property which is achieved during estimation through constrained optimization. Using notation from Bastin et al. (2010), the spline function which approximates the inverse cdf is expressed as

$$F_X^{-1}(u) \approx C(u) = \sum_{i=0}^{n} \pi_i N_{i,3}(u) \tag{4.13}$$

where n is the number of control points, π_i are the control points, and $N_{i,3}(u)$ are the cubic B-spline basis functions which depend on the knots. In general, complexity in spline functions is controlled by the number of knots and smoothness penalties (Ruppert et al., 2003, p 75). In practice, this approach has difficulty recovering the tails of the distribution, since spline functions are notoriously unreliable close to the boundaries of the knot vectors (Cirillo and Hetrakul, 2010).

Like the mixture of distributions MXL, the support of the sieve estimator with Legendre polynomials depends on the support of the base distribution used in the approximation. In contrast, the sieve estimator with B-splines assumes a bounded distribution, where the bounds are estimated from the data.

Besides boundedness, the two sieve estimators are alike in all other properties. They can both accommodate asymmetry and multiple modes. For the sieve estimator with Legendre polynomials, the number of modes is controlled by the length of the series expansion. For the sieve estimator with B-splines, the number of modes is controlled by the number of knots.

So far, neither sieve estimator can accommodate dependence between random parameters. The sieve estimators with Legendre polynomials and B-splines have only been developed and used for the univariate case. That is, they only estimate the distribution of each random parameter individually, implying mutual independence between random parameters. In order for these sieve estimators to accommodate dependence between random parameters, they would need to be extended to the multivariate case. Fosgerau (2014) describes another extension to accommodate dependence, by combining sieve estimators with copulas. Copulas impose a parametric form on the dependence structure, while the sieve estimator allows the marginal distributions to be arbitrary. However, the number and flexibility of multivariate copulas, and hence the flexibility of this extension, are limited.

4.5.3 Discrete or continuous?

Historically, the debate between continuous and discrete representations of heterogeneity has centred around the MXL with a parametric mixing distribution and the latent class logit model. The argument against the continuous representation is that it is parametric and therefore subject to misspecification, whereas the argument against the discrete approach is that it is too restrictive because it assumes homogeneity within each segment. Given the advanced mixing distributions which have been discussed in the previous two sections, these arguments are clearly outdated. The sieve estimators with distributions, Legendre polynomials and B-splines represent heterogeneity in a continuous manner while avoiding strict parametric assumptions. The fixed point MXL estimated on a fine grid of support points is a discrete representation which can capture far more detailed heterogeneity than any analytical distribution can display. What remains at the heart is the question: what is the true nature of heterogeneity, continuous or discrete?

This question is, practically speaking, unanswerable. At the same time, there are other, more relevant questions we could ask to determine whether to choose a continuous or discrete representation of heterogeneity. In reflecting on this debate, Wedel and Kamakura (2000, p 329) have the following to say:

> In applying models to segmentation, one should recognize that every model is at best a workable approximation of reality. One cannot claim that segments really exist or that the distributional form of unobserved heterogeneity is known.

The more relevant question we should ask ourselves is: which is the more *useful* representation, continuous or discrete? Wedel and Kamakura (2000) suggest that continuous representations are suitable for individual level forecasting, while discrete representations are particularly useful for understanding the structure of heterogeneity in the population. Segmentation is a core concept in marketing, not because it is necessarily more true than a continuous representation of heterogeneity, but because it has proven to be useful over and over again: it is accessible, compelling and actionable to managers and other end users of information resulting from marketing studies. On the other hand, when applied economists wish to generate population-level welfare estimates to be used in policymaking techniques such as cost-benefit analysis, segments are unnecessary, confusing and can lead to bias due to oversimplification from reducing a continuous distribution to a discrete distribution with a finite number of support points. If, however, policymakers are interested in the composition of 'winners' and 'losers' of a policy change, then segments once again become useful. To choose a continuous or a discrete representation, analysts should identify which levels of aggregation are appropriate for their context, research question and audience.

Continuous and discrete representations of heterogeneity can be complementary, each enriching the insight which can be gained from the other. In one example, Hynes et al. (2008) investigated preference heterogeneity among kayakers for whitewater sites in Ireland by estimating both a conventional MXL and a latent class logit model. With the conventional MXL, they established the presence of heterogeneity in the sample. With the latent class logit model, they were able to match the latent classes revealed by the model to specializations within the kayaking sport, with intuitive taste parameter estimates. For managers of the whitewater sites, this type of information is useful because it allows them to identify the mix of kayakers patronizing different sites and tailor responses to their preferences. In another example, Arunotayanun and Polak (2011) investigated mode choice heterogeneity among freight shippers. They used conventional MXL to establish the presence of heterogeneity even after the sample was split by commodity, the standard practice in this context, revealing the inadequacy of this segmentation scheme. They used the latent class logit model to identify alternative segments which were behaviourally driven instead, and systematically related them to shipper and shipment characteristics, leading to a new segmentation scheme. These examples illustrate how continuous and discrete mixing distributions can both be used to identify behaviourally and policy relevant heterogeneity.

For completeness, we bring to the reader's attention a nonparametric approach developed by Rouwendal et al. (2010) which does not rely on a statistical model and is not fully identified, but rather seeks to identify individual-specific valuations for attribute levels within a given dataset[9].

4.6 Empirical illustration of a nonparametric mixing distribution

In this section we offer empirical illustrations of discrete and continuous nonparametric mixing distributions.

[9] Rouwendal et al. (2010) introduces a method of locating individual-specific valuations of attributes by viewing each choice as revealing an inequality in valuations between attribute levels. For concreteness, we illustrate the core concept using a very simple example. Suppose that two bundles differ only in one attribute. If a respondent chooses bundle A over B, then he must have a higher valuation for the attribute level in bundle A than bundle B, thus revealing an inequality. Over the sequence of multiple choices, the space spanned by the inequalities shrinks, and ideally, is exactly identified. In an empirical application of the method, valuations were identified exactly in a few cases, and in most cases only bounds (which may be very wide) were identified. Furthermore, many respondents chose inconsistently, so that the space satisfying all inequalities was empty. Unlike the previously described approaches, this method avoids approximation and attempts to discover the specific valuations of each respondent in the dataset. It is not a statistical model which extends the valuations beyond the dataset at hand, and for most empirical cases the valuations will not be fully identified.

4.6.1 Discrete nonparametric mixing distribution

In section 4.2 we found strong empirical evidence for correlation between random parameters, which opens the question of whether the dependence structure is even more complex than mere correlation. Now that section 4.5 has introduced discrete nonparametric mixing distribution which can help answer that question, we can continue our line of empirical illustrations with the latent class logit model. We have already discussed how the LCL can be seen as the discrete analog to the commonly encountered mixed logit, with a continuous, parametric mixing distribution. Estimating the class-specific coefficients and the class membership probabilities is essentially estimating the location and mass of mass points in a pmf.

To maintain comparability with the previous alternative mixing distributions, only coefficients on the amount and location attributes were allowed to vary across classes. The other coefficients were constant across classes, resulting in a form of LCL sometimes called the 'equality constrained latent class' model in the attribute non-attendance literature (*e.g.* Scarpa et al., 2009; Hensher et al., 2012; Collins et al., 2013).

In order to maintain the discrete analogy to the baseline model, we fix class membership probabilities to be constant across individuals[10]. Thus, we estimate $k-1$ parameters, where k is the number of latent classes. Rather than estimating the membership probabilities directly, we estimate the parameters θ_c, which are used in a multinomial membership model as follows

$$\pi_c = \frac{\exp(\theta_c)}{\sum_k \exp(\theta_k)}$$

where π_c are the class membership probabilities. Estimating θ_c rather than the membership probabilities directly affords a computational advantage, because θ_c are unconstrained whereas the class membership probabilities must be proper probabilities.

Although the LCL is one of the most popular mixing distributions for the mixed logit model, the available software for it is scarce. For maximum flexibility, particularly in order to accommodate the 'equality constrained' nature of the LCL variation used in this empirical illustration, we found it expedient to code our own likelihood function and maximize it using the maxLik package (Henningsen and Toomet, 2011). We further found that starting values greatly affected the solution, a point also mentioned by Greene and Hensher (2003), so we used ten different starting values. The starting values were determined by estimating conditional logit models on a random partitioning of the dataset, as suggested by Train (2008).

[10]One advantage of the LCL is that it permits a class membership model, which can, for example, use sociodemographic variables to inform class assignment. However, this empirical illustration uses only the most basic class membership model.

One of the most important aspects of the LCL specification is the number of latent classes. This parameter is not estimated but rather user-specified. For many applications of the LCL, there is no theoretical guidance on the number of latent classes. Since there are also no statistical tests available to select the number of classes, the practitioner instead uses a combination of goodness-of-fit measures and individual discretion. Table 4.6 presents goodness-of-fit measures for the LCL estimated under one to six classes. The measures include the well-known information criteria AIC and BIC, as well as McFadden's pseudo R-squared, adjusted for degrees of freedom, and entropy (Morey et al., 2006). The entropy statistic is bounded between 0 and 1, and a value closer to 1 indicates that the model successfully differentiates individuals into latent classes. Note that a one-class LCL is equivalent to a standard conditional logit model, and that the 'infinite'-class LCL refers to the baseline model.

Number of classes	AIC	BIC	$\bar{\rho}^2$	Entropy
1	10260.52	10337.77	0.05	
2	9052.70	9155.70	0.16	0.16
3	8947.54	9076.29	0.17	0.22
4	8887.46	9041.96	0.18	0.13
5	8845.68	9025.93	0.18	0.14
6	8873.50	9079.50	0.18	0.14
∞	8947.67	9045.51	0.17	

Table 4.6: Goodness-of-fit measures for the LCL under different numbers of classes.

According to three of the four goodness-of-fit measures, the LCL with five latent classes has the best fit, even compared to the baseline model. Usually LCL models do not fit as well as a continuous mixture, so this result can be seen as evidence for the appropriateness of a discrete mixing distribution. On the basis of these goodness-of-fit measures and the plausibility of five classes, we select the five-class LCL for further comparison to the baseline model. Due to the number of estimates, the table of coefficient estimates is presented in section 6.G.

Reassuringly, the modes under the estimated LCL pmf and the estimated baseline pdf are very similar for each random parameter (Figure 4.9). However, the LCL pmf often reveals preference masses in the tails of the normal mixing distribution, thus biasing the mean and standard deviations of the LCL pmf with respect to the normal mixing distribution. Revealing these masses in the tails has both advantages and disadvantages. On the one hand, preferences not readily observable in the normal mixing distribution were revealed. On the other hand, these masses bias summary statistics like means and standard deviations. Thus, which mixing distribution is more appropriate depends on the policy motivation; if 'minority' preferences are policy relevant, then the LCL may be

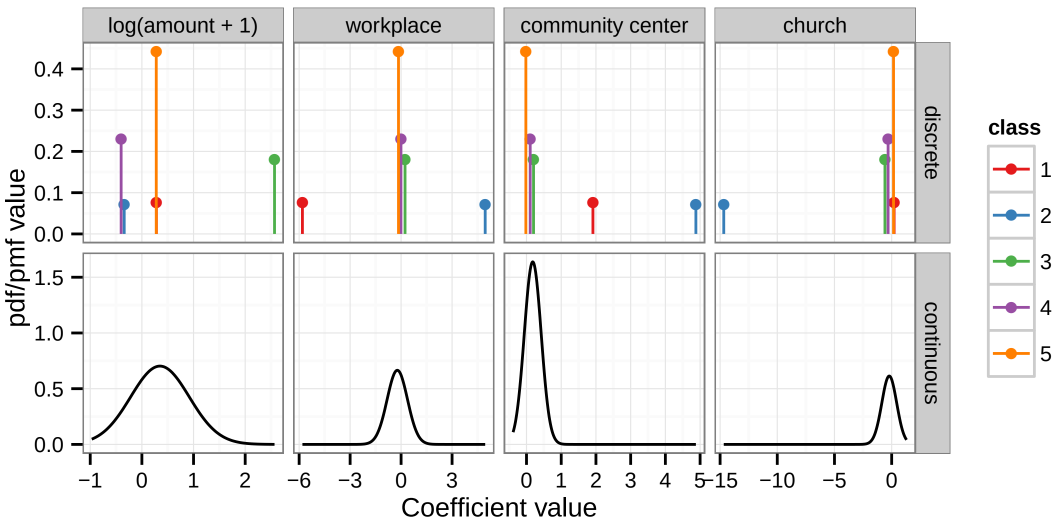


Figure 4.9: Discrete mixing distribution pmf and continuous mixing distribution pdf for each random parameter. The mass points are grouped by colours, representing the classes to which they belong. The colours and numbers associated with the classes are arbitrary.

more suitable. However, if the policy questions are primarily concerned with 'average' preferences across the whole population, then the normal mixing distribution may be a better, less biased choice.

One interesting approach to interpreting the LCL results is considering preference profiles by class, a capability not available under the baseline model (Figure 4.10). Class 5 captures the most individuals, with the highest membership probability, and is associated with preferences which are not extreme, although there is a clear positive preference for increasing reward amount. Classes 3 and 4 have the next highest membership probabilities, and are similar to class 5 with respect to location preferences, but have markedly different reward amount preferences. Class 4 has a negative preference for increasing reward amount while class 3 has a strongly positive preference for increasing reward amount. Classes 1 and 2, which have the lowest membership probabilities, have extreme preferences across location levels, with class 1 strongly preferring the community center over the workplace. Class 2, on the other hand, strongly prefers workplace and community center over the church. Since class 1 has a strong negative preference for workplace, we speculate that class 1 may consist of unemployed individuals[11]. The positive prefer-

[11] A speculation which could be checked by computing posterior class membership probabilities and examining the characteristics of individuals who have a high posterior probability for belonging to class

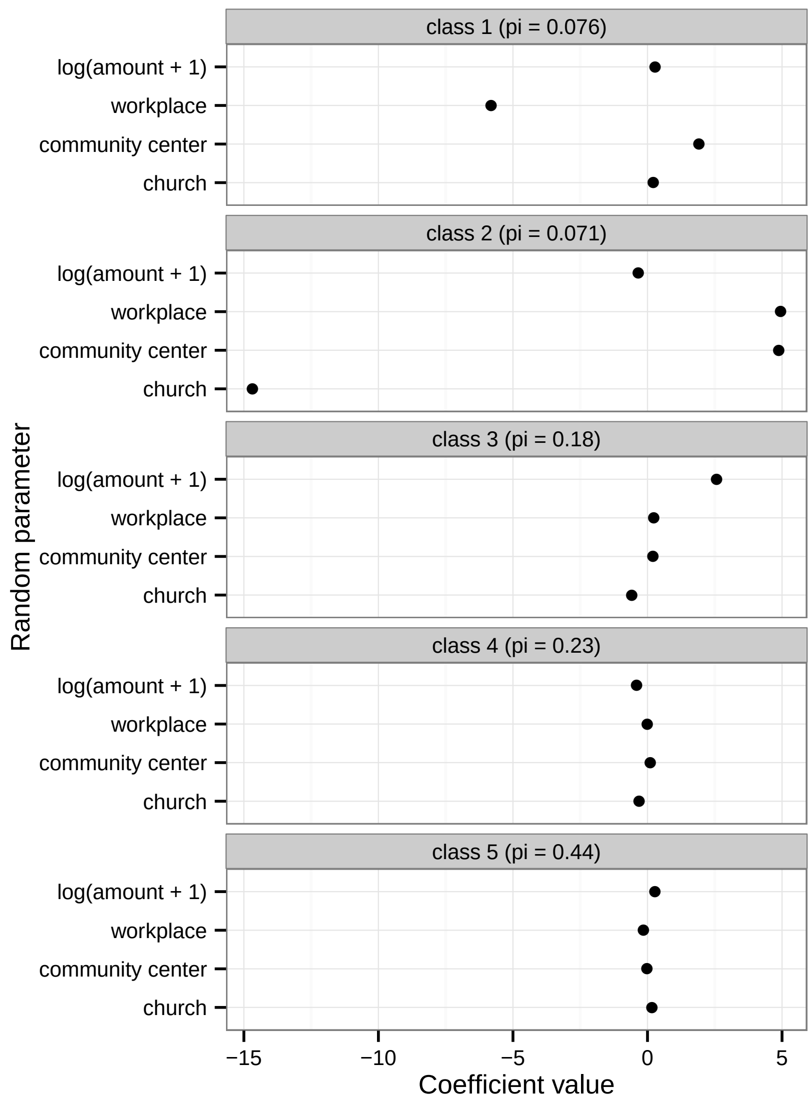


Figure 4.10: LCL coefficient estimates organized by class. Class membership probabilities are in parentheses.

ence for reward amount in class 1 is, thus, not so surprising, while the class 2 negative preference for increasing reward amount is more interesting.

This interpretation provides a richer picture of preferences than the baseline model in the event that market segments are policy relevant. In this example, the LCL suggests discrete market segments that could be targeted by obesity intervention providers. For example, a community center-based program which offers a strong financial incentive would attract classes 1 and 3, whereas a workplace-based program which offers a weak financial incentive would attract classes 2 and 4.

Moreover, the dependence between taste coefficients revealed by the latent classes cannot be captured by the normal mixing distribution, even if correlation were permitted. By estimating the joint distribution directly, the LCL permits arbitrary complexity in the dependence structure, including higher order relations beyond pairwise linear relations. For example, consider classes 1 and 2. Class 1 displays a negative relation between preferences for workplace and community center, while class 2 displays a positive relation (since both workplace and community center are preferred over church). Under a simple correlation matrix structure, we might estimate a zero correlation between workplace and community center due to the conflicting directions of these two classes. Moreover, the negative relation between workplace and community center is associated with a negative preference for increasing reward amount, while the positive relation in class 2 is associated with a positive preference for increasing reward.

Finally, we note that for three of the four random coefficients, the LCL pmf appears to be asymmetric. For example, the strong positive preference of class 3 for increasing reward amount is not matched by an equally strong negative preference. Similarly, the strongly negative preference of class 2 for the church location is not matched by an equally strong positive preference. However, the baseline mixing distribution is symmetric and cannot reveal the presence of asymmetry.

4.6.2 Continuous nonparametric mixing distribution

Although the evidence for asymmetry and multimodality we found in section 4.4 was weak, the weakness was due to lack of empirical identification. In this section we use a continuous nonparametric mixing distribution, a mixture of normals mixing distribution, in order to seek further evidence for asymmetry and multimodality.

Due to the lack of available software (a common issue with the nonparametric mixing distributions), we coded our own estimation routine based on the expectation-maximization algorithm introduced by Train (2008). Unfortunately, Train's EM algorithm

1.

has not yet been adapted for cases where some parameters remain fixed, and so we removed the fixed parameters from the model specification in this empirical illustration. Thus, the only parameters present in the specification are reward amount and program location. Although previous empirical applications of the EM algorithm for the mixture of normals distribution converged quickly (Train, 2008), our empirical application experienced very slow convergence and noisy log-likelihood paths over the iterations. Under the vast majority of starting values, the estimation failed to converge within 100 iterations.

As with the latent class logit model, the number of components is an important user-specified parameter. We estimated both two-component and three-component mixtures, and Table 4.7 suggests that the two-component mixture is more appropriate than the three-component mixture. However, the church coefficient displays very little heterogeneity compared to the baseline distribution (Figure 4.11), suggesting that this model may also suffer from the same weak empirical identification found when estimating the Johnson S_B and triangular mixing distributions. Consequently we also display the results from the three-component mixture, for which the variance of the coefficient distributions are all comparable to their respective baseline distributions (Figure 4.12). Table 4.7 also suggests that the baseline distribution is more appropriate than either of the alternative distributions. While overfitting is a possibility in this case, with the mixture of normals representing more complex heterogeneity than is actually present, the slow convergence of the EM algorithm could also be a culprit. If the algorithm were ended too early, then the optimal parameter estimates may not have been found.

	Log-likelihood	AIC	BIC
One component	−4454.833	8947.666	9071.47
Two components	−4473.473	8990.945	9304.617
Three components	−4469.097	9004.194	9474.702

Table 4.7: Goodness-of-fit statistics for mixture of normals distributions by number of components.

We observe that the means of both the two- and three-component mixtures tend to be shifted compared to the normal mixing distribution. This phenomenon may be attributed to omitted variable bias, since the remaining attributes after reward amount and program location were present in the baseline specification but not in the alternative specification.

Both figures also breaks down the alternative distribution into its component normal distributions (scaled by their shares). We can see that the components tend to vary primarily in their amount of variance rather than the locations of their means. Thus, the

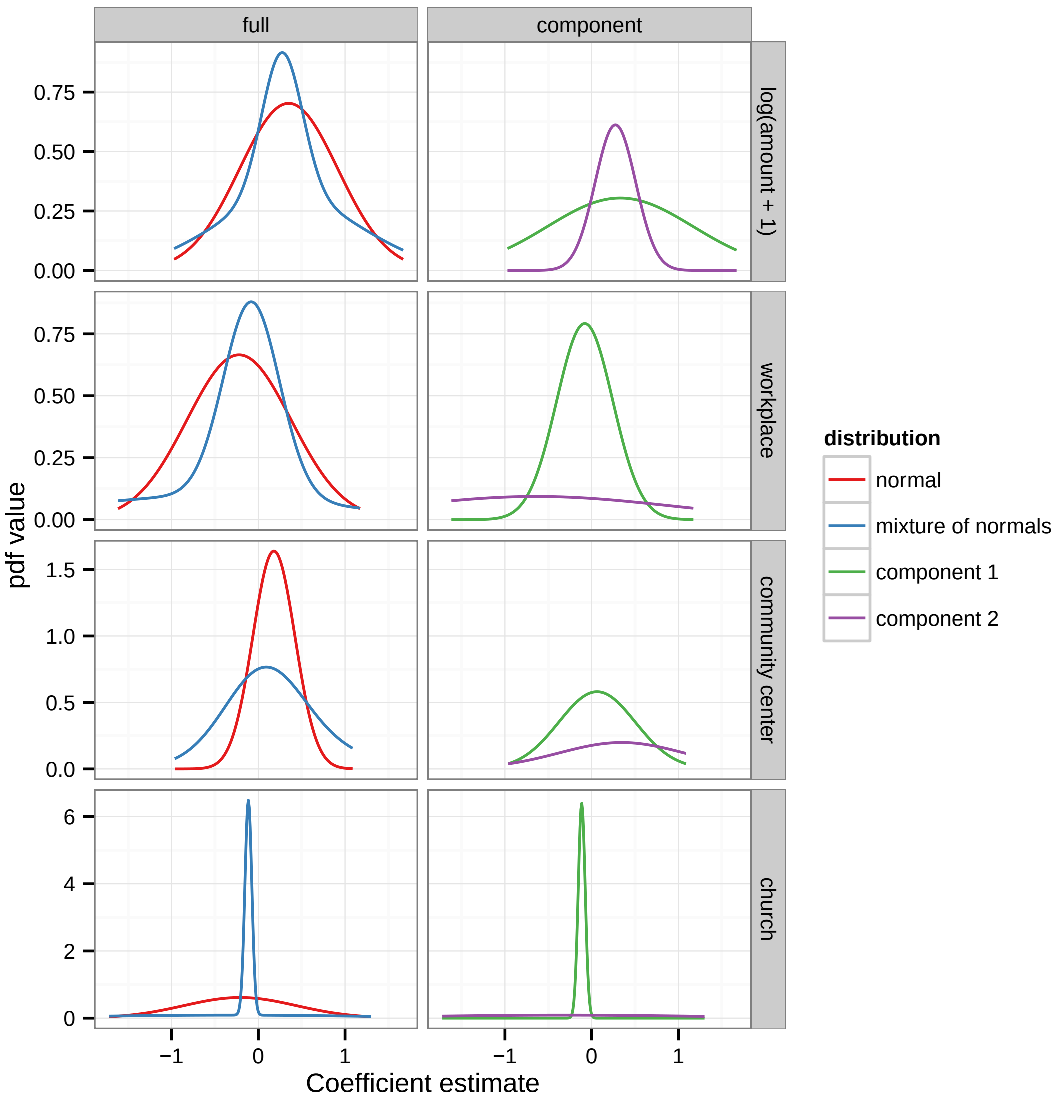


Figure 4.11: Density functions of two-component mixture and baseline distributions. Individual components of the mixture of normals distribution are shown on the right.

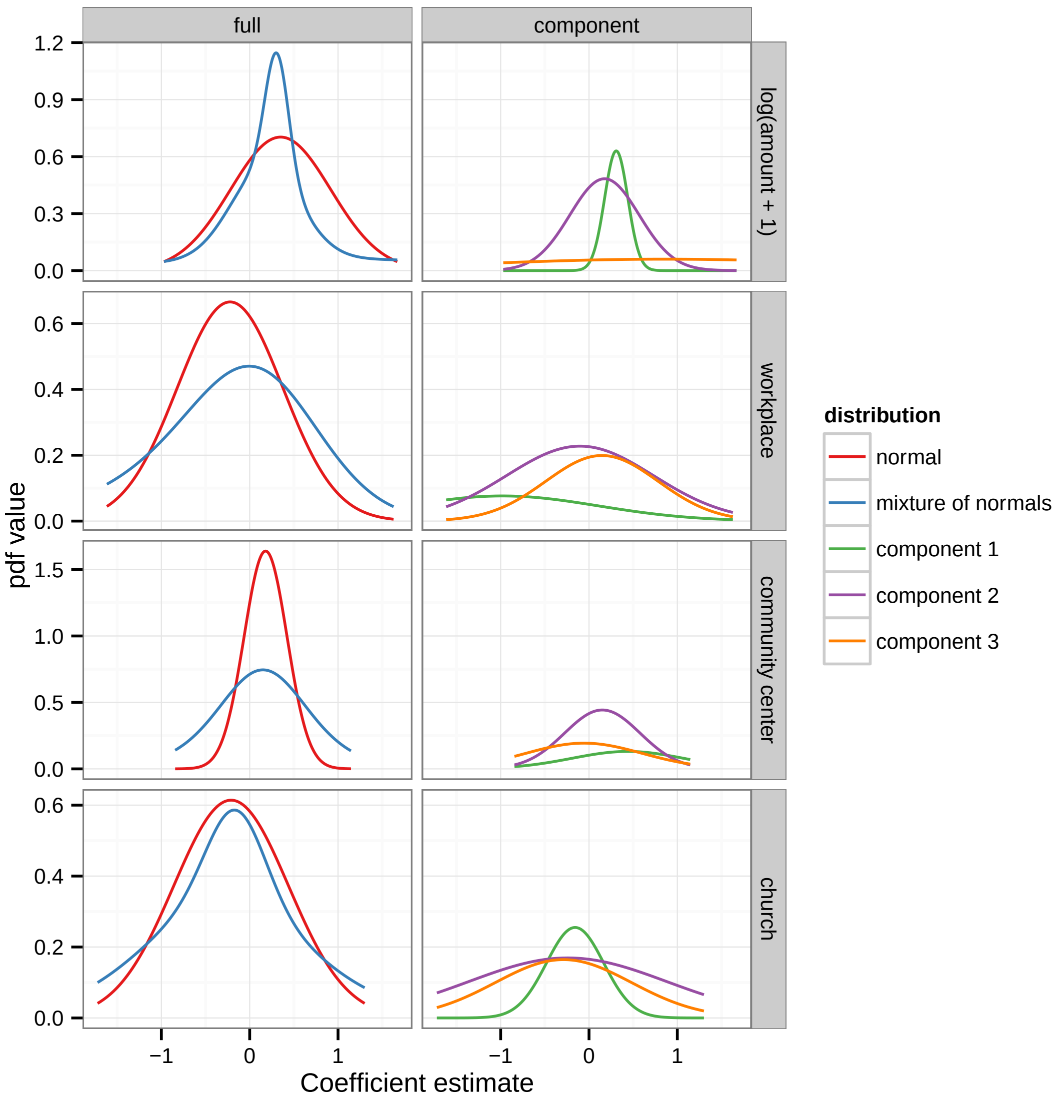


Figure 4.12: Density functions of three-component mixture and baseline distributions. Individual components of the mixture of normals distribution are shown on the right.

higher variance component tends to thicken the tails of the overall mixture, while the lower variance component provides the rest of the distribution shape.

This investigation revealed substantial evidence for asymmetry. In the case of a mixture of two normals, asymmetry occurs whenever the means of the two locations do not coincide. Thus, the relatively closely located means for the reward amount component distributions leads to a relatively symmetrical distribution. However, when a third component is brought into the mixture, asymmetry is much more evident. The mean component locations are much further apart for the location coefficients than the reward amount coefficient in the two-component mixture case, and consequently all the location coefficient distributions are skewed. In contrast, the community center coefficient distribution is relatively symmetrical in the three-component mixture, since the two lower-share components balance each other.

The left skew of the workplace and church coefficients, evident in both the two- and three-component mixtures, are not altogether surprising; if an individual is unemployed or doesn't attend church, then those program locations may arouse very negative preferences because they would exclude the individual from participation. The left skew in reward amount suggested by the three-component mixture lends support to the explanation that individuals offended by a financial reward for weight loss have strong negative preferences for increasing reward amount, unmatched by equally strong positive preferences from those who are not offended.

On the other hand, this investigation revealed no evidence for multimodality, supporting the results found in section 4.4. Unlike asymmetry, multimodality requires not only different mean locations, but also that the locations are far away enough, the variances small enough, or the shares equal enough that local modes are distinct. Since these conditions are not satisfied, a second or third distinct mode does not appear for any of the coefficients.

4.7 Parametric or nonparametric?

In Table 4.8, we summarise the key properties of each mixing distribution which have been discussed above. If flexibility and capturing as many features of heterogeneity as possible were the main goals of the analyst, then the nonparametric mixing distributions such as the fixed point and mixture of distributions MXL would be appropriate. However, more flexibility is not always better. Flexible mixing distributions may be overly complex, suffering from weak identification, overfitting and difficulty of interpretation. If the structure of the heterogeneity is simple, a simple mixing distribution would be more parsimonious, efficient, robust and interpretable (Keane and Wasi, 2013).

Mixing distribution	Support	Number of modes	Symmetry	Dependence
Parametric				
Normal	unbounded	1	symmetrical	correlation
Uniform	bounded	–	symmetrical	correlation
Triangular	bounded	1	a/symmetrical	correlation
Log-normal	bounded on one side	1	asymmetrical	correlation
Censored or truncated normal	bounded on one side	1	asymmetrical	correlation
Johnson S_B	bounded	2	a/symmetrical	correlation
Nonparametric discrete				
Latent class logit	bounded	n	a/symmetrical	arbitrary
Mass point	bounded	n	a/symmetrical	arbitrary
Fixed point	bounded	n	a/symmetrical	arbitrary
Nonparametric continuous				
Mixture of distributions	depends on component distributions	n	a/symmetrical	arbitrary
Sieve estimator with Legendre polynomials	depends on base distribution	n	a/symmetrical	independent
Sieve estimator with B-splines	bounded	n	a/symmetrical	independent

Table 4.8: Summary of properties by mixing distribution.

However, knowing whether heterogeneity is complex or simple prior to choosing a mixing distribution is difficult. Flexible mixing distributions can play a role here, by acting as a preliminary tool diagnosing which features of heterogeneity are present (Fosgerau and Hess, 2009)[12]. For example, if multiple modes are identified, then the Johnson S_B or nonparametric mixing distributions may be indicated. If a point mass at zero is identified, but the rest of the distribution appears smooth, then a censored normal or a latent class with fixed and random segments may be indicated. Campbell and Doherty (2013) implemented the latter approach to investigate WTP for value-added services to chicken. They hypothesized that some consumers would be indifferent to the value-added services, while others would have non-zero preferences. They accommodated these types by assigning each to a latent class: in one class, WTP was fixed to 0 for the former type of consumer, and in the other, WTP was allowed to be random. This approach allowed them to identify market niches, estimate demand responses to price premiums in those niches, and describe demographics of those niches which could aid in marketing efforts.

Using ten datasets, Keane and Wasi (2013) compared some of the mixing distributions described in this study, including the latent class logit model and the mixture of distributions approach. Although the emphasis of their study was on finding models of best fit, they found that the mixture of distributions model was superior to simpler mixing distributions in its ability to accommodate complex heterogeneity, in which small subpopulations of individuals have strong preferences for attributes to which most other individuals are indifferent. However, the mixture of distributions model accommodated these individuals by placing them into the tails of the component distributions. The latent class logit model, on the other hand, distinguished these individuals and placed them into their own segments. Even though the latent class logit model consistently scored among the lowest in terms of log-likelihood, it provided insight into the structure of heterogeneity hidden by other mixing distributions. Thus, more flexible distributions do not always provide more insight.

Practical considerations may also affect the choice of mixing distribution. In particular, mixing distributions vary in how easy they are to estimate, due to differences in

[12]Flexible mixing distributions are not the only such preliminary tool. Hensher and Greene (2003) propose an entirely different approach to 'reveal' the empirical distribution of heterogeneity. In this approach, a series of 'leave-one-out' MNL models are estimated, with each individual in the dataset 'left out' a different MNL model. The difference between the parameter estimate under a leave-one-out MNL model and under the pooled MNL model represents the individual's contribution to the mean parameter estimate. Considering all the differences from all of the leave-one-out MNL models presents a picture of individual heterogeneity. As a different approach, Hess (2010) suggests using conditional distributions (*i.e.*, respondent-specific coefficient distributions, conditional on the respondents' responses) to form a picture of the empirical distribution. However, he cautions against simply using the means of the conditional distributions, since doing so fails to take into account the heterogeneity around those means, within each conditional distribution. Conceptually, this approach is similar to checking one's prior with the estimated posterior in a Bayesian analysis.

empirical identifiability and available estimation strategies.

4.7.1 Estimation issues

Generally speaking, the more parameters a mixing distribution has, the more difficult it may be to empirically identify. For example, the Johnson S_B is a transformed normal with four parameters: two represent the mean and variance of the underlying normal and the other two represent the lower and upper bound of the distribution. Train and Sonnier (2005) suggests that the bounds are difficult to identify because the difference between them is closely related to the variance of the underlying normal. Consequently, Rigby et al. (2009) imposed bounds instead of allowing them to be freely estimated, Chalak et al. (2012) conducted a repetitive search for the bounds, and Cirillo and Hetrakul (2010) could not always achieve convergence during model estimation. The Johnson S_B is not the only parametric mixing distribution to suffer from estimation issues: the lognormal distribution occasionally does so as well (Hensher and Greene, 2003; Hess et al., 2006).

The nonparametric mixing distributions also vary in terms of empirical identification, but the degree to which they suffer does not seem to be closely related to the number of parameters involved. The latent class logit model, which is relatively parsimonious compared to, say, the fixed point MXL, often cannot be estimated beyond a handful of latent classes, as mentioned earlier. The mixture of distributions approach suffers from the same problem (*e.g.* Fosgerau and Hess, 2009), and few applications have attempted estimating more than two components in this model. The sieve estimators with Legendre polynomials and B-splines, on the other hand, have not presented analysts with any convergence problems. Similarly, the virtue and motivating reason to consider the fixed point MXL is its ability to remain fast and stable even under a huge number of points to estimate (Train, 2008).

Another factor affecting ease of estimation is what strategies are available for each mixing distribution. All of the continuous mixing distributions, parametric and nonparametric, are estimated by maximum simulated likelihood estimation (MSLE). The discrete mixing distributions, on the other hand, do not require simulation and can be estimated directly via maximum likelihood. Avoiding simulation is desirable because simulation increases computation time, introduces simulation bias, and requires a 'sufficient' number of draws in order to produce stable estimates. Unfortunately, determining this number is primarily an empirical matter (Hensher and Greene, 2003).

Train (2008) suggested using the expectation-maximization (EM) algorithm to avoid numerical problems which may be encountered during maximum likelihood estimation of models with a large number of parameters (*i.e.*, as typically found in nonparametric mixing distributions). He developed and illustrated the EM algorithm for several nonparametric mixing distributions, including the latent class logit model, the mixture of distributions (normals) MXL and the fixed point MXL. However, in a comparison of MSLE, the EM algorithm, and one other estimation strategy, Cherchi and Guevara (2012) found that as long as sample sizes were sufficient, MSLE outperformed the EM algorithm with respect to estimation time and estimation error. At the same time, as expected, MSLE suffered from weak empirical identification with insufficient sample sizes, while the EM algorithm was more robust to this issue. Other applications of the EM algorithm in the literature are limited, although Bastani and Weeks (2013) used it to estimate a fixed point MXL and latent class logit model, and Pacifico (2013) used it to estimate a latent class logit model. Neither study reported problems with using the EM algorithm.

4.8 Conclusion

In this study, we have described an extensive list of options for the choice of mixing distribution in a mixed logit model. Our descriptions focused on the theoretical ability of the distributions to describe heterogeneity features such as boundedness, multimodality, asymmetry and dependence between random parameters (summarised in Table 4.8), as well as discussing practical considerations such as estimation issues and the importance of the policy context in determining relevant heterogeneity features.

Each mixing distribution differs in its theoretical properties, practical considerations and policy context. The commonly proposed parametric mixing distributions typically have relatively inflexible theoretical properties, but can be easier to estimate and interpret. In many policy contexts, the additional flexibility in theoretical properties is unnecessary, because the policy relevant questions are at a high level, and the detailed features of the preference heterogeneity present do not affect the policy implications and may only complicate matters.

However, when the policy context calls for detailed consideration of the preference heterogeneity features, nonparametric mixing distributions which are more flexible than the parametric mixing distributions may be appropriate. When market segments are policy relevant, the latent class logit model can yield rich insights into minority preferences which would be glossed over in less flexible distributions, especially with respect to arbitrarily complex dependence structures. At the same time, population-wide conclusions based on the latent class logit model may give too much weight to the minority preferences. Moreover, in practice, the complexity of the latent class logit model is strongly limited by estimation issues which arise when too many latent classes are specified. This limitation has given rise to a number of variations on the latent class logit model, which also estimate a pmf for the mixing distribution. However, the way in which they parameterize the support points is different, and so fewer hyperparameters are necessary to estimate the same number of support points. One weakness common to all the discrete nonparametric mixing distributions is the large extent to which results depend on analyst-specified tuning parameters, such as number and range of support points.

The continuous nonparametric mixing distributions are more forgiving, with results fairly robust to tuning parameters such as number and choice of basis function. These mixing distributions are all sieve estimators, which approximate unknown functions with a series of basis functions. They can also be more parsimonious than the discrete nonparametric mixing distributions in terms of numbers of estimated hyperparameters, while still maintaining a high degree of flexibility. Their chief weakness is their inability, at present, to accommodate dependence between random coefficients. The discrete nonparametric mixing distributions, in contrast, directly estimate the joint mixing distribution, and dependence between random coefficients is revealed as a by-product.

In our empirical comparison of the baseline mixing distribution, a multivariate normal with mutually independent random parameters, and alternative distributions which relax the theoretical properties of the baseline distribution, we found the biggest differences were in dependence structure. When we compared the baseline distribution with a correlated normal mixing distribution, we found significant correlation between different levels within the same attribute. Furthermore, when we compared the baseline distribution with the latent class logit model, we found evidence of dependence relationships which were more complex than the pairwise, linear relationships modelled by the correlated normal. The market segments identified by the latent class logit model suggested different designs to attract different segments.

In contrast, our empirical comparison of the baseline distribution with a bounded alternative, the uniform distribution, did not reveal substantial differences in policy implication. Although technically there were significant differences in standard deviation between the baseline and uniform distributions, these differences did not have much effect on the policy implications. The amount of the normal density found outside of the uniform density was modest, and evenly distributed between the upper and lower tails. This even distribution can be attributed to the symmetrical nature of both distributions and the similarity in their locations. Consequently, considering only a specific portion of the distribution, such as the negative preferences, again showed very small differences between the baseline and alternative distributions. Thus, unless the extreme values of the distribution (*e.g.*, top or bottom 1%) are policy relevant, there is little difference in policy implications between unbounded and bounded distributions.

Our other empirical comparisons with alternative parametric mixing distributions were less revealing, in part due to weak empirical identification. However, we were able to find more conclusive evidence for asymmetry by using a continuous nonparametric distribution, the mixture of normals mixing distribution. We found skew in the workplace and church coefficients, which could be explained by the exclusive nature of those venues for individuals who are unemployed and do not attend church. We also found some evidence of negative skew in the reward amount coefficient in the three-component mixture, the Johnson S_B and the triangular distribution.

We found essentially no evidence for multimodality in any of the coefficient distributions. Although the Johnson S_B distribution. The Johnson S_B distribution suggested that the reward amount coefficient was unimodal, but its inflexibility in representing bimodal distributions (U-shape only) may have made identifying bimodality or multimodality difficult. Using, instead, the mixture of normals distribution, we found no evidence of multimodality.

The weak empirical identification we experienced suggest more robust implementations of maximum simulated likelihood routines for the triangular and Johnson S_B distributions are called for. The algorithmic implementation of the maximum simulated likelihood approach for uncommon mixing distributions is a weakness of our study and potential future direction for future study. In general, the lack of readily available software to estimate mixed logit models with alternative mixing distributions is an obstacle to increasing the adoption of these distributions. Thus, robust, fast, and user-friendly implementations of alternative mixing distributions are much needed.

When analysts have a choice of which heterogeneity features to model, it makes sense to start with a more flexible mixing distribution and then simplify as features are seen to be absent. Although we discussed this role flexible mixing distributions could play in section 4.7, doing so is not yet common practice. Future work could focus on this role, establishing which of the flexible distributions are most appropriate for aiding the specification search.

The importance of dependence structure, particularly moving beyond correlation into higher-order relationships, suggests that approaches for capturing these relationships is an important future direction for theoretical study. Given the large amount of data needed to estimate the joint pmf directly in nonparametric discrete mixing distributions, approaches which are elegant and parsimonious could be complementary to the existing techniques for detecting complex dependence structure. However, given the limitations of the copula approach and the difficulty with estimating highly multidimensional nonparametric surfaces, this direction appears to be quite challenging.

Bibliography

Allenby, Greg M. and Rossi, Peter E. Marketing models of consumer heterogeneity. *Journal of Econometrics*, 89(1–2):57–78, 1998. ISSN 0304-4076. doi: 10.1016/S0304-4076(98)00055-4. URL http://www.sciencedirect.com/science/article/pii/S0304407698000554.

Amador, Francisco Javier, González, Rosa Marina, and de Dios Ortúzar, Juan. On confounding preference heterogeneity and income effect in discrete choice models. *Networks and Spatial Economics*, 8(2-3):97–108, 2008. ISSN 1566-113X. doi: 10.1007/s11067-007-9043-6. URL http://dx.doi.org/10.1007/s11067-007-9043-6.

Arunotayanun, Kriangkrai and Polak, John W. Taste heterogeneity and market segmentation in freight shippers' mode choice behaviour. *Transportation Research Part E: Logistics and Transportation Review*, 47(2):138 – 148, 2011. ISSN 1366-5545. doi: http://dx.doi.org/10.1016/j.tre.2010.09.003. URL http://www.sciencedirect.com/science/article/pii/S1366554510000827.

Bajari, Patrick, Fox, Jeremy T., and Ryan, Stephen P. Linear regression estimation of discrete choice models with nonparametric distributions of random coefficients. *The American Economic Review*, 97(2):pp. 459–463, 2007. ISSN 00028282. URL http://www.jstor.org/stable/30034495.

Balcombe, Kelvin, Burton, Michael, and Rigby, Dan. Skew and attribute non-attendance within the bayesian mixed logit model. *Journal of Environmental Economics and Management*, 62(3):446 – 461, 2011. ISSN 0095-0696. doi: http://dx.doi.org/10.1016/j.jeem.2011.04.004. URL http://www.sciencedirect.com/science/article/pii/S0095069611000635.

Bastani, Parisa and Weeks, Melvyn. Nonparametric estimation of mixing distributions in discrete choice models: Heterogeneity and WTP. In *European Economic Association & Econometric Society Parallel Meetings*, Gothenburg, August 2013. URL http://www.eea-esem.com/eea-esem/2013/prog/viewpaper.asp?pid=2232.

Bastin, Fabian, Cirillo, Cinzia, and Toint, Philippe L. Estimating nonparametric random utility models with an application to the value of time in heterogeneous populations. *Transportation Science*, 44(4):537–549, 2010. doi: 10.1287/trsc.1100.0321. URL http://dx.doi.org/10.1287/trsc.1100.0321.

Bierens, Herman J. Semi-nonparametric interval-censored mixed proportional hazard models: Identification and consistency results. *Econometric Theory*, 24:749–794, 6 2008. ISSN 1469-4360. doi: 10.1017/S0266466608080316. URL http://journals.cambridge.org/article_S0266466608080316.

Birol, Ekin, Karousakis, Katia, and Koundouri, Phoebe. Using a choice experiment to account for preference heterogeneity in wetland attributes: The case of cheimaditida wetland in greece. *Ecological Economics*, 60(1):145 – 156, 2006. ISSN 0921-8009. doi: http://dx.doi.org/10.1016/j.ecolecon.2006.06.002. URL http://www.sciencedirect.com/science/article/pii/S0921800906002989.

Bujosa, Angel, Riera, Antoni, and Hicks, Robert L. Combining discrete and continuous representations of preference heterogeneity: A latent class approach. *Environmental and Resource Economics*, 47(4):477–493, 2010. ISSN 0924-6460. doi: 10.1007/s10640-010-9389-y. URL http://dx.doi.org/10.1007/s10640-010-9389-y.

Campbell, Danny and Doherty, Edel. Combining discrete and continuous mixing distributions to identify niche markets for food. *European Review of Agricultural Economics*, 40(2):287–312, 2013. doi: 10.1093/erae/jbs018. URL http://erae.oxfordjournals.org/content/40/2/287.abstract.

Campbell, Danny, Hess, Stephane, Scarpa, Riccardo, and Rose, John M. Using finite mixture models to accommodate outliers in discrete choice modelling. In *International Choice Modelling Conference*, 2009. URL http://www.icmconference.org.uk/index.php/icmc/icmc2009/paper/view/105.

Campbell, Danny, Doherty, Edel, Hynes, Stephen, and van Rensburg, Tom M. Combining discrete and continuous mixing approaches to accommodate heterogeneity in price sensitivities in environmental choice analysis. In *84th Annual Conference*, Edinburgh, Scotland, March 29-31 2010. Agricultural Economics Society. URL http://purl.umn.edu/91723.

Campbell, Danny, Vedel, Suzanne Elizabeth, Thorsen, Bo Jellesmark, and Jacobsen, Jette Bredahl. Heterogeneity in the WTP for recreational access: distributional aspects. *Journal of Environmental Planning and Management*, 57(8):1200–1219, 2014. doi: 10.1080/09640568.2013.793173. URL http://dx.doi.org/10.1080/09640568.2013.793173.

Chalak, Ali, Hecht, Jonathan, Reid, Scott, and Abiad, Mohamad G. Willingness-to-pay for greenhouse gas reductions: A Bayesian investigation of distributional patterns. *Environmental Science & Policy*, 19–20(0):147–157, 2012. ISSN 1462-9011. doi: http://dx.doi.org/10.1016/j.envsci.2012.02.003. URL http://www.sciencedirect.com/science/article/pii/S1462901112000330.

Cherchi, Elisabetta and de Dios Ortúzar, Juan. Can mixed logit reveal the actual data generating process? some implications for environmental assessment. *Transportation Research Part D: Transport and Environment*, 15(7):428 – 442, 2010. ISSN 1361-9209. doi: http://dx.doi.org/10.1016/j.trd.2010.04.010. URL http://www.sciencedirect.

com/science/article/pii/S1361920910000702. Specification and interpretation issues in behavioural models used for environmental assessment.

Cherchi, Elisabetta and Guevara, Cristian Angelo. A Monte Carlo experiment to analyze the curse of dimensionality in estimating random coefficients models with a full variance-covariance matrix. *Transportation Research Part B: Methodological*, 46(2):321–332, 2012. ISSN 0191-2615. doi: http://dx.doi.org/10.1016/j.trb.2011.10.006. URL http://www.sciencedirect.com/science/article/pii/S0191261511001469. Emerging and Innovative Directions in Choice Modeling.

Cherchi, Elisabetta and Ortúzar, Juan de Dios. *Predicting best with mixed logit models: understanding some confounding effects*, chapter 6, pages 215–235. Nova Science Publishers, Inc., 2008.

Cirillo, Cinzia and Hetrakul, Pratt. Continuous random coefficient logit models: A comparison of parametric and non-parametric methods to estimate individual preferences over cybernetic transportation systems. In *Transportation Research Board 89th Annual Meeting*, 2010. URL http://trid.trb.org/view/2010/C/909334.

Collins, Andrew T., Rose, John M., and Hensher, David A. Specification issues in a generalised random parameters attribute nonattendance model. *Transportation Research Part B: Methodological*, 56(0):234 – 253, 2013. ISSN 0191-2615. doi: http://dx.doi.org/10.1016/j.trb.2013.08.001. URL http://www.sciencedirect.com/science/article/pii/S0191261513001318.

Croissant, Yves. *mlogit: multinomial logit model*, 2013. URL http://CRAN.R-project.org/package=mlogit. R package version 0.2-4.

Daly, Andrew, Hess, Stephane, and Train, Kenneth. Assuring finite moments for willingness to pay in random coefficient models. *Transportation*, 39(1):19–31, 2012. ISSN 0049-4488. doi: 10.1007/s11116-011-9331-3. URL http://dx.doi.org/10.1007/s11116-011-9331-3.

Dekker, Thijs. The asymmetric triangular distribution for continuous mixed multinomial logit models. In *Transportation Research Board 93rd Annual Meeting*, 2014. URL http://trid.trb.org/view.aspx?id=1288992.

Desarbo, Wayne, Ansari, Asim, Chintagunta, Pradeep, Himmelberg, Charles, Jedidi, Kamel, Johnson, Richard, Kamakura, Wagner, Lenk, Peter, Srinivasan, Kannan, and Wedel, Michel. Representing heterogeneity in consumer response models 1996 choice conference participants. *Marketing Letters*, 8(3):335–348, 1997. ISSN 0923-0645. doi: 10.1023/A:1007916714911. URL http://dx.doi.org/10.1023/A%3A1007916714911.

Dong, X. and Koppelman, F. S. Comparison of continuous and discrete representations of unobserved heterogeneity. *J Market Anal*, 2(1):43–58, Mar 2014. ISSN 2050-3318. URL http://dx.doi.org/10.1057/jma.2014.5.

Dong, Xiaojing and Koppelman, Frank S. Comparison of methods representing heterogeneity in logit models. In *10th International Conference on Travel Behaviour Research*, Lucerne, 2003. URL https://edit.ethz.ch/ivt/news/archive/20030810_IATBR/dong.pdf.

Fosgerau, Mogens. *Handbook of choice modelling*, chapter Nonparametric approaches to describing heterogeneity, pages 257–267. Edward Elgar Publishing, 2014. doi: 10.4337/9781781003152.00018. URL http://www.elgaronline.com/view/9781781003145.00018.xml.

Fosgerau, Mogens and Bierlaire, Michel. A practical test for the choice of mixing distribution in discrete choice models. *Transportation Research Part B: Methodological*, 41 (7):784 – 794, 2007. ISSN 0191-2615. doi: http://dx.doi.org/10.1016/j.trb.2007.01.002. URL http://www.sciencedirect.com/science/article/pii/S0191261507000136.

Fosgerau, Mogens and Hess, Stephane. A comparison of methods for representing random taste heterogeneity in discrete choice models. *European Transport*, 42:1–25, 2009. URL http://www.openstarts.units.it/dspace/bitstream/10077/6127/1/42_1_FosgerauHess.pdf.

Greene, William H. and Hensher, David A. A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research Part B: Methodological*, 37(8):681 – 698, 2003. ISSN 0191-2615. doi: http://dx.doi.org/10.1016/S0191-2615(02)00046-2. URL http://www.sciencedirect.com/science/article/pii/S0191261502000462.

Greene, William H. and Hensher, David A. Revealing additional dimensions of preference heterogeneity in a latent class mixed multinomial logit model. *Applied Economics*, 45(14):1897–1902, 2013. doi: 10.1080/00036846.2011.650325. URL http://www.tandfonline.com/doi/abs/10.1080/00036846.2011.650325.

Halpern, Scott D., Madison, Kristin M., and Volpp, Kevin G. Patients as mercenaries?: The ethics of using financial incentives in the war on unhealthy behaviors. *Circulation: Cardiovascular Quality and Outcomes*, 2(5):514–516, 2009. doi: 10.1161/CIRCOUTCOMES.109.871855. URL http://circoutcomes.ahajournals.org/content/2/5/514.short.

Henningsen, Arne and Toomet, Ott. maxlik: A package for maximum likelihood estimation in R. *Computational Statistics*, 26(3):443–458, 2011. doi: 10.1007/s00180-010-0217-1. URL http://dx.doi.org/10.1007/s00180-010-0217-1.

Hensher, David A. and Greene, William H. The mixed logit model: The state of practice. *Transportation*, 30(2):133–176, 2003. ISSN 0049-4488. doi: 10.1023/A:1022558715350. URL http://dx.doi.org/10.1023/A%3A1022558715350.

Hensher, DavidA., Rose, JohnM., and Greene, WilliamH. Inferring attribute non-attendance from stated choice data: implications for willingness to pay estimates and a warning for stated choice experiment design. *Transportation*, 39(2):235–245, 2012. ISSN 0049-4488. doi: 10.1007/s11116-011-9347-8. URL http://dx.doi.org/10.1007/s11116-011-9347-8.

Hess, S., Bierlaire, M., and Polak, J. W. A systematic comparison of continuous and discrete mixture models. *European Trans*, 8(37):35–51, 2007. URL http://hdl.handle.net/10077/5957.

Hess, Stephane. Conditional parameter estimates from mixed logit models: distributional assumptions and a free software tool. *Journal of Choice Modelling*, 3(2):134–152, 2010. URL http://www.sciencedirect.com/science/article/pii/S1755534513700393.

Hess, Stephane. *Handbook of Choice Modelling*, chapter Latent class structures: taste heterogeneity and beyond, pages 311–330. Edward Elgar Publishing, 2014. doi: 10.4337/9781781003152.00021. URL http://www.elgaronline.com/view/9781781003145.00021.xml.

Hess, Stephane and Rose, JohnM. Can scale and coefficient heterogeneity be separated in random coefficients models? *Transportation*, 39(6):1225–1239, 2012. ISSN 0049-4488. doi: 10.1007/s11116-012-9394-9. URL http://dx.doi.org/10.1007/s11116-012-9394-9.

Hess, Stephane, Bierlaire, Michel, and Polak, John W. Estimation of value of travel-time savings using mixed logit models. *Transportation Research Part A: Policy and Practice*, 39 (2Ű3):221 – 236, 2005. ISSN 0965-8564. doi: http://dx.doi.org/10.1016/j.tra.2004.09.007. URL http://www.sciencedirect.com/science/article/pii/S0965856404001028. Positive Utility of Travel Positive Utility of Travel.

Hess, Stephane, Axhausen, Kay W., and Polak, John W. Distributional assumptions in mixed logit. In *Transportation Research Board 85th Annual Meeting*, 2006. URL http://trid.trb.org/view.aspx?id=777238.

Hess, Stephane, Ben-Akiva, Moshe, and Gopinath, Dinesh. Advantages of latent class over continuous mixture of logit models. Working paper, May 2011.

Horsky, Dan, Misra, Sanjog, and Nelson, Paul. Observed and unobserved preference heterogeneity in brand-choice models. *Marketing Science*, 25(4):322–335, 2006. doi: 10.

1287/mksc.1050.0192. URL http://pubsonline.informs.org/doi/abs/10.1287/mksc.1050.0192.

Hynes, Stephen, Hanley, Nick, and Scarpa, Riccardo. Effects on welfare measures of alternative means of accounting for preference heterogeneity in recreational demand models. *American Journal of Agricultural Economics*, 90(4):1011–1027, 2008. doi: 10.1111/j.1467-8276.2008.01148.x. URL http://ajae.oxfordjournals.org/content/90/4/1011.abstract.

Johnson, N. L. Systems of frequency curves generated by methods of translation. *Biometrika*, 36(1/2):pp. 149–176, 1949. ISSN 00063444. URL http://www.jstor.org/stable/2332539.

Johnson, N. L. and Kitchen, J. O. Some notes on tables to facilitate fitting sb curves. *Biometrika*, 58(1):pp. 223–226, 1971. ISSN 00063444. URL http://www.jstor.org/stable/2334335.

Johnston, Robert J. and Duke, Joshua M. Willingness to pay for agricultural land preservation and policy process attributes: Does the method matter? *American Journal of Agricultural Economics*, 89(4):1098–1115, 2007. doi: 10.1111/j.1467-8276.2007.01029.x. URL http://ajae.oxfordjournals.org/content/89/4/1098.abstract.

Keane, Michael and Wasi, Nada. Comparing alternative models of heterogeneity in consumer choice behavior. *Journal of Applied Econometrics*, 28(6):1018–1045, 2013. ISSN 1099-1255. doi: 10.1002/jae.2304. URL http://dx.doi.org/10.1002/jae.2304.

Lunze, Karsten and Paasche-Orlow, Michael K. Financial incentives for healthy behavior: Ethical safeguards for behavioral economics. *American Journal of Preventive Medicine*, 44 (6):659 – 665, 2013. ISSN 0749-3797. doi: http://dx.doi.org/10.1016/j.amepre.2013.01.035. URL http://www.sciencedirect.com/science/article/pii/S0749379713001864.

Mabit, Stefan, Hess, Stephane, and Cassaude, Sebastiàn. Correlation and willingness-to-pay indicators in transport demand modelling. In *Transportation Research Board Annual Meeting*, 2008. URL http://trid.trb.org/view.aspx?id=847647.

Marcucci, Edoardo and Gatta, Valerio. Dissecting preference heterogeneity in consumer stated choices. *Transportation Research Part E: Logistics and Transportation Review*, 48(1): 331 – 339, 2012. ISSN 1366-5545. doi: http://dx.doi.org/10.1016/j.tre.2011.08.003. URL http://www.sciencedirect.com/science/article/pii/S1366554511001086. Select Papers from the 19th International Symposium on Transportation and Traffic Theory.

McFadden, Daniel and Train, Kenneth. Mixed MNL models for discrete response. *Journal of Applied Econometrics*, 15(5):447–470, 2000. ISSN 1099-1255. doi: 10.1002/

1099-1255(200009/10)15:5<447::AID-JAE570>3.0.CO;2-1. URL http://dx.doi.org/10.1002/1099-1255(200009/10)15:5<447::AID-JAE570>3.0.CO;2-1.

Morey, Edward, Thacher, Jennifer, and Breffle, William. Using angler characteristics and attitudinal data to identify environmental preference classes: A latent-class model. *Environmental and Resource Economics*, 34(1):91–115, 2006. ISSN 0924-6460. doi: 10.1007/s10640-005-3794-7. URL http://dx.doi.org/10.1007/s10640-005-3794-7.

Pacifico, Daniele. On the role of unobserved preference heterogeneity in discrete choice models of labour supply. *Empirical Economics*, 45(2):929–963, 2013. ISSN 0377-7332. doi: 10.1007/s00181-012-0637-6. URL http://dx.doi.org/10.1007/s00181-012-0637-6.

Provencher, Bill and Bishop, Richard C. Does accounting for preference heterogeneity improve the forecasting of a random utility model? a case study. *Journal of Environmental Economics and Management*, 48(1):793 – 810, 2004. ISSN 0095-0696. doi: http://dx.doi.org/10.1016/j.jeem.2003.11.001. URL http://www.sciencedirect.com/science/article/pii/S0095069603001359.

Provencher, Bill and Moore, Rebecca. A discussion of "using angler characteristics and attitudinal data to identify environmental preference classes: A latent-class model". *Environmental and Resource Economics*, 34(1):117–124, 2006. ISSN 0924-6460. doi: 10.1007/s10640-005-3793-8. URL http://dx.doi.org/10.1007/s10640-005-3793-8.

Provencher, Bill, Baerenklau, Kenneth A., and Bishop, Richard C. A finite mixture logit model of recreational angling with serially correlated random utility. *American Journal of Agricultural Economics*, 84(4):1066–1075, 2002. doi: 10.1111/1467-8276.00054. URL http://ajae.oxfordjournals.org/content/84/4/1066.abstract.

Rigby, Dan, Balcombe, Kelvin, and Burton, Michael. Mixed logit model performance and distributional assumptions: Preferences and gm foods. *Environmental and Resource Economics*, 42(3):279–295, 2009. ISSN 0924-6460. doi: 10.1007/s10640-008-9227-7. URL http://dx.doi.org/10.1007/s10640-008-9227-7.

Roth, Alvin E. Repugnance as a constraint on markets. *Journal of Economic Perspectives*, 21(3):37–58, 2007. doi: 10.1257/jep.21.3.37. URL http://www.aeaweb.org/articles.php?doi=10.1257/jep.21.3.37.

Rouwendal, Jan, de Blaeij, Arianne, Rietveld, Piet, and Verhoef, Erik. The information content of a stated choice experiment: A new method and its application to the value of a statistical life. *Transportation Research Part B: Methodological*, 44(1):136 – 151, 2010. ISSN 0191-2615. doi: http://dx.doi.org/10.1016/j.trb.2009.04.006. URL http://www.sciencedirect.com/science/article/pii/S0191261509000599.

Ruppert, David, Wand, Matt P, and Carroll, Raymond J. *Semiparametric regression*. Number 12. Cambridge university press, 2003.

Scarpa, Riccardo, Thiene, Mara, and Marangon, Francesco. Using flexible taste distributions to value collective reputation for environmentally friendly production methods. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie*, 56 (2):145–162, 2008. ISSN 1744-7976. doi: 10.1111/j.1744-7976.2008.00122.x. URL http://dx.doi.org/10.1111/j.1744-7976.2008.00122.x.

Scarpa, Riccardo, Gilbride, Timothy J., Campbell, Danny, and Hensher, David A. Modelling attribute non-attendance in choice experiments for rural landscape valuation. *European Review of Agricultural Economics*, 36(2):151–174, 2009. doi: 10.1093/erae/jbp012. URL http://erae.oxfordjournals.org/content/36/2/151.abstract.

Schmidt, Harald. Wellness incentives, equity, and the 5 groups problem. *Am J Public Health*, 102(1):49–54, November 2011. ISSN 0090-0036. doi: 10.2105/AJPH.2011.300348. URL http://dx.doi.org/10.2105/AJPH.2011.300348.

Sillano, M and de Ortúzar, J D. Willingness-to-pay estimation with mixed logit models: some new evidence. *Environment and Planning A*, 37(3):525–550, 2005. URL http://www.envplan.com/abstract.cgi?id=a36137.

Stein, William E. and Keblis, Matthew F. A new method to simulate the triangular distribution. *Mathematical and Computer Modelling*, 49(5-6):1143 – 1147, 2009. ISSN 0895-7177. doi: http://dx.doi.org/10.1016/j.mcm.2008.06.013. URL http://www.sciencedirect.com/science/article/pii/S0895717708002665.

Train, Kenneth. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2009.

Train, Kenneth and Sonnier, Garrett. *Mixed Logit with Bounded Distributions of Correlated Partworths*, volume 6 of *The Economics of Non-Market Goods and Resources*, pages 117–134. Springer Netherlands, 2005. ISBN 978-1-4020-3683-5. doi: 10.1007/1-4020-3684-1_7. URL http://dx.doi.org/10.1007/1-4020-3684-1_7.

Train, Kenneth E. EM algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modelling*, 1(1):40–69, 2008. ISSN 1755-5345. doi: http://dx.doi.org/10.1016/S1755-5345(13)70022-8. URL http://www.sciencedirect.com/science/article/pii/S1755534513700228.

Wedel, Michel and Kamakura, Wagner. *Market segmentation: conceptual and methodological foundations*. Kluwer Academic Publishers, second edition edition, 2000.

Wedel, Michel, Kamakura, Wagner, Arora, Neeraj, Bemmaor, Albert, Chiang, Jeongwen, Elrod, Terry, Johnson, Rich, Lenk, Peter, Neslin, Scott, and Poulsen, CarstenStig. Discrete and continuous representations of unobserved heterogeneity in choice modeling. *Marketing Letters*, 10(3):219–232, 1999. ISSN 0923-0645. doi: 10.1023/A:1008054316179. URL http://dx.doi.org/10.1023/A%3A1008054316179.

Chapter 5

Conclusion

This dissertation considered different forms of preference heterogeneity in discrete choice models through three essays on a case study in financial incentives for weight loss programs. The first essay used a model with conditional heterogeneity to target demographics vulnerable to overweight and obesity or understudied in the weight loss literature. The second essay investigated the ability of a polytomous attendance scale to elicit more reliable self-reports of attribute attendance behaviour, specific case of decision strategy heterogeneity. The third essay explored available mixing distributions to model features of unobserved heterogeneity such as boundedness, modality, symmetry and dependence between random coefficients.

The first essay modelled preference heterogeneity conditional on observed demographics, so that demographic-specific preferences were revealed. These demographic-specific preferences were then used to identify financial incentive designs which would be most attractive to target demographics vulnerable to overweight and obesity or understudied in the weight loss literature. We found three main policy implications, 1) customizing the incentive design can enhance participation, 2) large reward amounts can actually be counterproductive to participation and closing health disparities, and 3) a better strategy for increasing participation is to offer multiple programs with different incentive designs. The first implication follows from the preference heterogeneity across demographic groups, which implies that the groups are differentiated in their preferences. Hence, custom programs can be designed to match each preference profile. The second implication follows from the existence of a segment of the population which has a negative response to increasing reward amount, a preference particularly prevalent in the African American subpopulation. In this case, more is not better, because a higher reward amount alienates individuals in an target demographic group, thereby failing to reduce health disparities. The third implication follows from the preference heterogeneity both across and within demographic groups. To attract more target demographic groups and more individuals within each demographic, the offered programs could be made heterogeneous to match the preference heterogeneity evident. The diminishing marginal returns in participation when offering more programs suggest that offering two programs affords the greatest return on participation.

For the analyst, future work could focus on gaining a more nuanced understanding of 1) preferences by demographic group, 2) the negative response to reward amount and 3) program location. For the first direction, in the current analysis, demographic groups are not mutually exclusive. Although the obese group and African American group have, in some senses, 'opposite' preferences, there may actually be considerable overlap between the two groups. Separating out preference heterogeneity by weight status and demographic group (as well as comparing low-income women to higher income women, for example), could provide a more nuanced understanding of preferences by demographic group. For the second direction, previous literature on repugnance in economic transactions has shown that such transactions can be conducted with more acceptance under certain conditions. Since the negative response to reward amount is counterproductive to the policy goal of enhancing participation, future work could focus understanding the basis for the negative response and strategies for ameliorating it, such as the moderating influence payment form could have on reward amount. For the third direction, program location

For the practitioner, future work could examine alternate approaches to increasing participation and representativeness, instead of or in addition to financial incentives. For example, different intervention modalities, such as Internet delivery, could increase participation by affording greater convenience. The result that program location is one of the most attributes suggests that a different modality which enhances attractiveness in terms of location could result in considerable gains.

The first essay found that reward amount and program location were by far the most important attributes to respondents, and that reward frequency was the least important attribute to most respondents. Since some of the attributes were less important than others, these findings suggest the possibility for attribute non-attendance. The second essay explored that possibility using polytomous attendance scales.

The second essay modelled decision strategy heterogeneity conditional on self-reported attendance behaviour. Self-reported attendance behaviour had been found to be unreliable in previous studies, but this study used a more flexible, polytomous attendance scale which had the potential to address the sources of unreliability. The main implication of this essay is that the nature of information revealed by attribute attendance statements is still unclear, in spite of using a polytomous attendance scales. This implication follows from the finding that few coefficients accorded well with theoretical expectations, such as significant differences between attendance levels and greater sensitivity with higher attendance levels for preferred attribute levels. Inconsistency was also found when comparing to inferred attendance, with many more dissimilar classifications of attendance state than similar classifications. However, we did find evidence that the lowest two points on the polytomous attendance scale came closest to the assumption of zero marginal utility usually associated with attribute non-attendance.

For the analyst, future work could focus on measuring attribute attendance behaviour using alternate approaches which focus on behaviours which are less than fully conscious (*e.g.*, eye tracking). However, the attendance information revealed by conscious and subconscious/unconscious methods may be complementary rather than redundant, suggesting further work in understanding what type of information is revealed by explicit attendance statements, if it is not zero marginal utility as previously expected.

For the practitioner, the recommendation is to be aware of the potential for attribute non-attendance, and assess it on a case-by-case basis considering the policy context. In this case of financial incentives for weight loss programs, there is a similar level of potential compared to many choice experiments, in which the choice task is unfamiliar because the participants have rarely or never considered the choice at hand. However, the low levels of stated non-attendance and the mixed results from including attendance information suggest that the policy implications from the first essay are not substantively impacted by attribute non-attendance, in spite of the potential for biased estimates.

In the first two essays, we used observable markers of heterogeneity such as demographic variables and attendance statements in order to model heterogeneity. However, there will always be some amount of heterogeneity left that does not have any other explanation. The third essay focused on this last part of heterogeneity, unobserved heterogeneity, through the selection of mixing distributions for the mixed logit model.

The third essay recommended selecting mixing distributions on the basis of heterogeneity features which are relevant to policy and theory, such as range of the support, number of modes, symmetry and dependence between random parameters. The essay then described parametric, discrete nonparametric and continuous nonparametric distributions in terms of which heterogeneity features each could capture. The main implication of this discussion is that the choice of a mixing distribution involves trade-offs between flexibility and stability. Parametric distributions are elegant and stable in estimation, but their very simplicity results in strong restrictions on the heterogeneity features which can be captured. Discrete nonparametric distributions are more flexible in their ability to capture many of the heterogeneity features discussed, but can be highly unstable in estimation. Continuous nonparametric distributions represent a compromise between parametric and discrete nonparametric mixing distributions, providing greater flexibility with fewer hyperparameters, but their ability to capture complex dependence between random parameters is limited. These findings suggest that future work for analysts could be to improve the stability and tuning parameter selection for discrete nonparametric distributions and develop multivariate approaches for continuous nonparametric distributions.

The third essay also illustrated the discussed heterogeneity features empirically using a selection of parametric and nonparametric mixing distributions. We found that a bounded support did not substantially change the policy implications as compared to an unbounded support. We also found strong evidence for correlation between random parameters, and some evidence for more complex dependence relations as well as the presence of asymmetry. We did not find any evidence for multiple modes. Some alternative distributions were subject to weak empirical identification, suggesting two paths for future work: 1) overfitting and 2) software. Empirical identification could be weak due to overfitting, which would occur if there is insufficient heterogeneity in the taste parameters. However, the significant heterogeneity we found in most taste parameters under the baseline mixing distribution suggests that the amount of heterogeneity is adequate, and so what is lacking may be the complexity of the heterogeneity structure. Future work could address the trade-off between useful evidence and overly complex representations of heterogeneity. Poor implementation of the estimation algorithm could also be responsible for weak empirical identification. Software for alternative mixing distributions is not widely available, and this lack can be a serious obstacle to using the appropriate mixing distribution. Thus, making stable and fast algorithms for a number of mixing distributions more widely available is another path for future work.

It is interesting to contrast the first and third essays. In the first essay, the structure of preference heterogeneity is imposed externally, using observed demographics. In the third essay, the structure of preference heterogeneity is revealed internally, since the heterogeneity is driven by choice behaviour rather than a covariate. The first essay examined relative preference strength across taste parameters, while the third essay examined relative preference strength within each taste parameter. Thus, the results from the two essays are not directly comparable although they both consider preference heterogeneity. What would tie the results from the two essays together is a broader theory of heterogeneity, as described by Wedel et al. (1999).

However, in the absence of such a theory, what drives the specification of heterogeneity is rather policy goal. In the first essay, the policy goal was to target demographic groups which have been identified as being vulnerable to overweight and obesity or understudied in the financial incentives for weight loss literature. Hence, the heterogeneity was structured in terms of those target demographic groups. The third essay considered other policy needs which could be answered by different heterogeneity features. To highlight the difference, consider the concept of minority preferences. The first essay was concerned with minority preferences: each of the target demographic groups was a minority of the national population. However, the 'minority' in this case was defined externally. The third essay also identified minority preferences, using the latent class logit model. However, in this case 'minority' was internally defined, because it was revealed by choice behaviour. This approach could answer the policy goal of reaching the most participants in the national population, by catering to the majority preferences. This goal is in contrast to the first essay's goal, which was to reach participants in *a priori* defined groups. It is important to note that the two minorities may not consist of the same individuals at all. To identify individuals in the internally defined minorities, it would be necessary to conduct post-estimation analysis, such as computing the respondent-specific conditional coefficient distributions.

The broader theory of heterogeneity is also needed to ensure that unobserved heterogeneity represents taste variation and not other omitted structures. Modelling unobserved heterogeneity without taking into account any other forms of preference heterogeneity can be misleading because the hyperparameters describing unobserved heterogeneity tend to absorb any specification errors such as additional forms of heterogeneity (Provencher and Bishop, 2004; Amador et al., 2008). Thus it is not always clear whether the heterogeneity seen in model estimates is true taste variation or some other type of heterogeneity. For analysts, future work could be in the direction of developing such a broad theory. Practitioners can contribute to this direction by conducting more empirical applications, adding to a body of work which could result in empirical generalizations as called for by Wedel et al. (1999).

Bibliography

Amador, Francisco Javier, González, Rosa Marina, and de Dios Ortúzar, Juan. On confounding preference heterogeneity and income effect in discrete choice models. *Networks and Spatial Economics*, 8(2-3):97–108, 2008. ISSN 1566-113X. doi: 10.1007/s11067-007-9043-6. URL http://dx.doi.org/10.1007/s11067-007-9043-6.

Provencher, Bill and Bishop, Richard C. Does accounting for preference heterogeneity improve the forecasting of a random utility model? a case study. *Journal of Environmental Economics and Management*, 48(1):793 – 810, 2004. ISSN 0095-0696. doi: http://dx.doi.org/10.1016/j.jeem.2003.11.001. URL http://www.sciencedirect.com/science/article/pii/S0095069603001359.

Wedel, Michel, Kamakura, Wagner, Arora, Neeraj, Bemmaor, Albert, Chiang, Jeongwen, Elrod, Terry, Johnson, Rich, Lenk, Peter, Neslin, Scott, and Poulsen, CarstenStig. Discrete and continuous representations of unobserved heterogeneity in choice modeling. *Marketing Letters*, 10(3):219–232, 1999. ISSN 0923-0645. doi: 10.1023/A:1008054316179. URL http://dx.doi.org/10.1023/A%3A1008054316179.

Chapter 6

Appendices

6.A Primary conditional logit model estimates

	Coefficient estimates
ASC.SQ	-0.32 $(0.15)^{*}$
‘log(Mag + 1)‘	0.21 $(0.06)^{***}$
Mag.0	0.06 (0.16)
LocationWorkplace	-0.10 (0.07)
LocationCommunity center	0.19 $(0.07)^{**}$
LocationChurch	-0.25 $(0.07)^{***}$
Typegym	-0.00 (0.08)
Typemedical	-0.20 $(0.07)^{**}$
Typedebit	0.27 $(0.07)^{***}$
Contingentweight	0.02 (0.08)
Contingentcompliance	-0.05 (0.08)
Contingentatt.comp	-0.07 (0.08)
Nrewardsweekly	0.09 (0.08)
Nrewardsmonthly	0.02 (0.08)
Nrewardsquarterly	0.07 (0.07)
‘race.ethnicity1:Mag.0‘	0.08 (0.16)
‘race.ethnicity2:Mag.0‘	0.14 (0.16)
‘race.ethnicity3:Mag.0‘	-0.09 (0.27)
‘femaleTRUE:Mag.0‘	-0.09 (0.09)
‘income1:Mag.0‘	0.28 (0.21)
‘income2:Mag.0‘	-0.15 (0.15)
‘income3:Mag.0‘	-0.35 $(0.15)^{*}$

	Coefficient estimates
‘employedTRUE:Mag.0‘	−0.02 (0.09)
‘education1:Mag.0‘	0.28 (0.13)*
‘education2:Mag.0‘	0.11 (0.12)
‘bmi.status1:Mag.0‘	0.09 (0.22)
‘bmi.status2:Mag.0‘	0.09 (0.14)
‘race.ethnicity1:log(Mag + 1)‘	0.01 (0.06)
‘race.ethnicity2:log(Mag + 1)‘	−0.09 (0.06)
‘race.ethnicity3:log(Mag + 1)‘	0.09 (0.10)
‘femaleTRUE:log(Mag + 1)‘	0.00 (0.03)
‘income1:log(Mag + 1)‘	−0.08 (0.07)
‘income2:log(Mag + 1)‘	0.10 (0.05)
‘income3:log(Mag + 1)‘	0.07 (0.05)
‘employedTRUE:log(Mag + 1)‘	−0.05 (0.03)
‘education1:log(Mag + 1)‘	−0.14 (0.05)**
‘education2:log(Mag + 1)‘	−0.02 (0.04)
‘bmi.status1:log(Mag + 1)‘	−0.10 (0.08)
‘bmi.status2:log(Mag + 1)‘	−0.03 (0.05)
‘race.ethnicity1:LocationWorkplace‘	0.06 (0.07)
‘race.ethnicity2:LocationWorkplace‘	−0.08 (0.07)
‘race.ethnicity3:LocationWorkplace‘	0.03 (0.11)
‘race.ethnicity1:LocationCommunity center‘	−0.10 (0.07)
‘race.ethnicity2:LocationCommunity center‘	−0.01 (0.07)
‘race.ethnicity3:LocationCommunity center‘	0.10 (0.11)
‘race.ethnicity1:LocationChurch‘	0.12 (0.07)
‘race.ethnicity2:LocationChurch‘	0.02 (0.07)
‘race.ethnicity3:LocationChurch‘	−0.18 (0.12)
‘femaleTRUE:LocationWorkplace‘	−0.01 (0.04)
‘femaleTRUE:LocationCommunity center‘	0.04 (0.04)
‘femaleTRUE:LocationChurch‘	−0.11 (0.04)*
‘income1:LocationWorkplace‘	−0.17 (0.10)
‘income2:LocationWorkplace‘	−0.07 (0.07)
‘income3:LocationWorkplace‘	0.08 (0.07)
‘income1:LocationCommunity center‘	0.04 (0.09)
‘income2:LocationCommunity center‘	0.04 (0.07)
‘income3:LocationCommunity center‘	0.04 (0.07)
‘income1:LocationChurch‘	0.05 (0.09)
‘income2:LocationChurch‘	0.06 (0.07)
‘income3:LocationChurch‘	−0.05 (0.07)

	Coefficient estimates
‘employedTRUE:LocationWorkplace‘	-0.25 (0.04)***
‘employedTRUE:LocationCommunity center‘	0.08 (0.04)
‘employedTRUE:LocationChurch‘	0.10 (0.04)*
‘education1:LocationWorkplace‘	0.12 (0.06)
‘education2:LocationWorkplace‘	-0.11 (0.05)
‘education1:LocationCommunity center‘	-0.12 (0.06)*
‘education2:LocationCommunity center‘	0.14 (0.05)**
‘education1:LocationChurch‘	0.03 (0.06)
‘education2:LocationChurch‘	-0.04 (0.05)
‘bmi.status1:LocationWorkplace‘	0.21 (0.10)*
‘bmi.status2:LocationWorkplace‘	-0.07 (0.07)
‘bmi.status1:LocationCommunity center‘	0.08 (0.10)
‘bmi.status2:LocationCommunity center‘	-0.02 (0.06)
‘bmi.status1:LocationChurch‘	-0.15 (0.10)
‘bmi.status2:LocationChurch‘	0.10 (0.07)
‘race.ethnicity1:Typegym‘	-0.10 (0.08)
‘race.ethnicity2:Typegym‘	0.14 (0.08)
‘race.ethnicity3:Typegym‘	0.10 (0.12)
‘race.ethnicity1:Typemedical‘	0.05 (0.07)
‘race.ethnicity2:Typemedical‘	0.01 (0.08)
‘race.ethnicity3:Typemedical‘	-0.05 (0.12)
‘race.ethnicity1:Typedebit‘	0.03 (0.07)
‘race.ethnicity2:Typedebit‘	-0.16 (0.07)*
‘race.ethnicity3:Typedebit‘	0.06 (0.12)
‘femaleTRUE:Typegym‘	-0.07 (0.05)
‘femaleTRUE:Typemedical‘	0.04 (0.04)
‘femaleTRUE:Typedebit‘	0.07 (0.04)
‘income1:Typegym‘	-0.10 (0.10)
‘income2:Typegym‘	-0.16 (0.07)*
‘income3:Typegym‘	-0.04 (0.07)
‘income1:Typemedical‘	-0.01 (0.10)
‘income2:Typemedical‘	0.07 (0.07)
‘income3:Typemedical‘	-0.02 (0.07)
‘income1:Typedebit‘	-0.00 (0.10)
‘income2:Typedebit‘	-0.00 (0.07)
‘income3:Typedebit‘	0.05 (0.07)
‘employedTRUE:Typegym‘	0.07 (0.05)
‘employedTRUE:Typemedical‘	-0.02 (0.04)

	Coefficient estimates
‘employedTRUE:Typedebit‘	−0.05 (0.04)
‘education1:Typegym‘	0.15 (0.07)*
‘education2:Typegym‘	−0.01 (0.06)
‘education1:Typemedical‘	−0.11 (0.06)
‘education2:Typemedical‘	0.11 (0.06)
‘education1:Typedebit‘	0.07 (0.06)
‘education2:Typedebit‘	−0.01 (0.06)
‘bmi.status1:Typegym‘	0.04 (0.11)
‘bmi.status2:Typegym‘	−0.02 (0.07)
‘bmi.status1:Typemedical‘	−0.03 (0.11)
‘bmi.status2:Typemedical‘	−0.01 (0.07)
‘bmi.status1:Typedebit‘	0.16 (0.11)
‘bmi.status2:Typedebit‘	0.00 (0.07)
‘race.ethnicity1:Contingentweight‘	0.03 (0.07)
‘race.ethnicity2:Contingentweight‘	0.14 (0.08)
‘race.ethnicity3:Contingentweight‘	−0.19 (0.13)
‘race.ethnicity1:Contingentcompliance‘	−0.04 (0.07)
‘race.ethnicity2:Contingentcompliance‘	−0.05 (0.08)
‘race.ethnicity3:Contingentcompliance‘	0.06 (0.12)
‘race.ethnicity1:Contingentatt.comp‘	−0.10 (0.07)
‘race.ethnicity2:Contingentatt.comp‘	−0.07 (0.07)
‘race.ethnicity3:Contingentatt.comp‘	−0.03 (0.11)
‘femaleTRUE:Contingentweight‘	0.07 (0.05)
‘femaleTRUE:Contingentcompliance‘	0.08 (0.04)
‘femaleTRUE:Contingentatt.comp‘	−0.13 (0.04)**
‘income1:Contingentweight‘	0.06 (0.10)
‘income2:Contingentweight‘	0.02 (0.07)
‘income3:Contingentweight‘	−0.05 (0.07)
‘income1:Contingentcompliance‘	0.25 (0.10)**
‘income2:Contingentcompliance‘	0.11 (0.07)
‘income3:Contingentcompliance‘	−0.05 (0.07)
‘income1:Contingentatt.comp‘	−0.17 (0.09)
‘income2:Contingentatt.comp‘	−0.05 (0.07)
‘income3:Contingentatt.comp‘	0.03 (0.07)
‘employedTRUE:Contingentweight‘	−0.12 (0.05)**
‘employedTRUE:Contingentcompliance‘	0.05 (0.04)
‘employedTRUE:Contingentatt.comp‘	0.05 (0.04)
‘education1:Contingentweight‘	0.11 (0.06)

	Coefficient estimates
‘education2:Contingentweight‘	-0.04 (0.06)
‘education1:Contingentcompliance‘	-0.09 (0.06)
‘education2:Contingentcompliance‘	-0.06 (0.06)
‘education1:Contingentatt.comp‘	0.06 (0.06)
‘education2:Contingentatt.comp‘	0.06 (0.06)
‘bmi.status1:Contingentweight‘	-0.05 (0.11)
‘bmi.status2:Contingentweight‘	0.02 (0.07)
‘bmi.status1:Contingentcompliance‘	0.02 (0.11)
‘bmi.status2:Contingentcompliance‘	-0.03 (0.07)
‘bmi.status1:Contingentatt.comp‘	-0.05 (0.12)
‘bmi.status2:Contingentatt.comp‘	0.03 (0.07)
‘race.ethnicity1:Nrewardsweekly‘	-0.01 (0.08)
‘race.ethnicity2:Nrewardsweekly‘	-0.05 (0.08)
‘race.ethnicity3:Nrewardsweekly‘	0.20 (0.13)
‘race.ethnicity1:Nrewardsmonthly‘	-0.03 (0.07)
‘race.ethnicity2:Nrewardsmonthly‘	-0.05 (0.08)
‘race.ethnicity3:Nrewardsmonthly‘	-0.02 (0.12)
‘race.ethnicity1:Nrewardsquarterly‘	0.01 (0.07)
‘race.ethnicity2:Nrewardsquarterly‘	-0.03 (0.08)
‘race.ethnicity3:Nrewardsquarterly‘	0.08 (0.12)
‘femaleTRUE:Nrewardsweekly‘	-0.00 (0.05)
‘femaleTRUE:Nrewardsmonthly‘	0.06 (0.04)
‘femaleTRUE:Nrewardsquarterly‘	-0.02 (0.04)
‘income1:Nrewardsweekly‘	0.11 (0.11)
‘income2:Nrewardsweekly‘	0.05 (0.08)
‘income3:Nrewardsweekly‘	-0.13 (0.08)
‘income1:Nrewardsmonthly‘	-0.04 (0.10)
‘income2:Nrewardsmonthly‘	-0.01 (0.07)
‘income3:Nrewardsmonthly‘	0.02 (0.07)
‘income1:Nrewardsquarterly‘	-0.04 (0.10)
‘income2:Nrewardsquarterly‘	0.04 (0.07)
‘income3:Nrewardsquarterly‘	-0.01 (0.07)
‘employedTRUE:Nrewardsweekly‘	-0.07 (0.05)
‘employedTRUE:Nrewardsmonthly‘	0.00 (0.04)
‘employedTRUE:Nrewardsquarterly‘	0.01 (0.04)
‘education1:Nrewardsweekly‘	-0.16 (0.07)*
‘education2:Nrewardsweekly‘	0.06 (0.06)
‘education1:Nrewardsmonthly‘	0.14 (0.06)*

	Coefficient estimates
'education2:Nrewardsmonthly'	−0.04 (0.06)
'education1:Nrewardsquarterly'	0.01 (0.06)
'education2:Nrewardsquarterly'	−0.01 (0.06)
'bmi.status1:Nrewardsweekly'	0.05 (0.12)
'bmi.status2:Nrewardsweekly'	−0.10 (0.08)
'bmi.status1:Nrewardsmonthly'	0.10 (0.11)
'bmi.status2:Nrewardsmonthly'	−0.08 (0.07)
'bmi.status1:Nrewardsquarterly'	−0.03 (0.11)
'bmi.status2:Nrewardsquarterly'	0.08 (0.07)
race.ethnicity1	−0.07 (0.15)
race.ethnicity2	0.25 (0.16)
race.ethnicity3	−0.13 (0.25)
femaleTRUE	−0.14 (0.09)
income1	0.19 (0.20)
income2	−0.21 (0.14)
income3	−0.13 (0.14)
employedTRUE	−0.01 (0.09)
education1	0.22 (0.13)
education2	0.06 (0.11)
bmi.status1	0.01 (0.21)
bmi.status2	0.09 (0.14)
Log-likelihood	−4392.02
N	4572.00

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

6.B Determining the functional form of reward amount

Lancsar and Louviere (2008) recommend visualizing implied functional forms for attribute variables by estimating coefficients on each attribute level. When we employ this technique on the reward amount attribute, we find that treating reward amount as a categorical variable suggests a logarithmic transformation for reward amount (Figure 6.1).

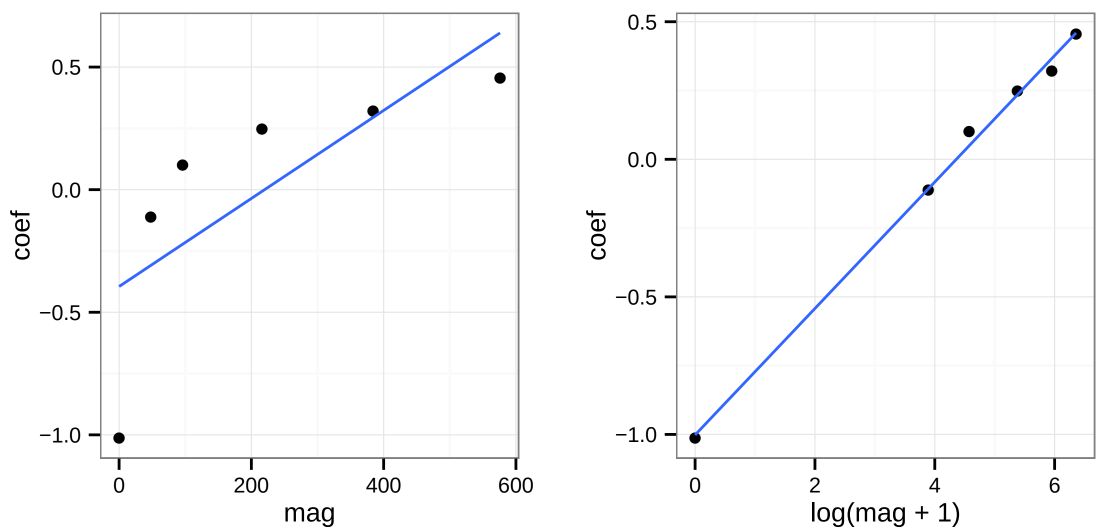


Figure 6.1: Coefficient estimates for reward amount as a categorical variable. The left-hand panel displays the estimates for reward amount levels without transformation, and the right-hand panel displays the estimates for log-transformed reward amount levels.

6.C Recovering effects on interactions between effect-coded variables

Many choice modelling studies use dummy coded categorical variables. Dummy coding provides simple contrasts, in which coefficient estimates for a given factor level are interpreted in comparison to a reference level, which is omitted from the model matrix. Unfortunately, the ommitted levels are then perfectly collinear with intercept in the systematic utility specification, so preferences for the omitted levels cannot be recovered.

To avoid this problem, several authors recommend using effect coding instead (Holmes and Adamowicz, 2003; Hoyos, 2010; Bech and Gyrd-Hansen, 2005). Effect coding provides deviation contrasts, in which coefficient estimates for a given factor level are interpreted in comparison to the grand mean for all factor levels. Consequently, the coefficient for the omitted level, while not directly estimated, can be recovered by taking the negative sum of the coefficient estimates for the other levels in the factor.

To make this discussion concrete, consider the following simple example: suppose there is a categorical variable x_1 with three levels, $\{l_1^1, l_2^1, l_3^1\}$, where the third level is the omitted level. Under dummy coding, we would code x_1 with two variables, x_1^1 and x_2^1, as follows:

x_1	x_1^1	x_2^1
l_1^1	1	0
l_2^1	0	1
l_3^1	0	0

Under effect coding, we would instead code x_1 as follows:

x_1	x_1^1	x_2^1
l_1^1	1	0
l_2^1	0	1
l_3^1	-1	-1

and the coefficient on the omitted level, l_3^1 could be recovered as the negative sum of the coefficients on the other two levels, $-(b_1^1 + b_2^1)$.

The preceding is all common knowledge, but applies only to effect-coded variables entering into a specification on their own. For effect-coded variables which are interacted, the situation is somewhat more complicated. However, coefficients on omitted interactions (*i.e.*, interactions involving omitted levels) can still be recovered by taking the appropriate sums of coefficients. To illustrate how to arrive at the appropriate sum, consider the following simple example. Suppose we have, in addition to the categorical variables x_1 with three levels mentioned above, another categorical variable x_2, also with three levels, and the third level as the omitted level. The contrast matrix is now much larger:

x_1	x_2	x_1^1	x_2^1	x_1^2	x_2^2	$x_1^1 \times x_1^2$	$x_2^1 \times x_1^2$	$x_1^1 \times x_2^2$	$x_2^1 \times x_2^2$
l_1^1	l_1^2	1	0	1	0	1	0	0	0
l_2^1	l_1^2	0	1	1	0	0	1	0	0
l_3^1	l_1^2	-1	-1	1	0	-1	-1	0	0
l_1^1	l_2^2	1	0	0	1	0	0	1	0
l_2^1	l_2^2	0	1	0	1	0	0	0	1
l_3^1	l_2^2	-1	-1	0	1	0	0	-1	-1
l_1^1	l_3^2	1	0	-1	-1	-1	0	-1	0
l_2^1	l_3^2	0	1	-1	-1	0	-1	0	-1
l_3^1	l_3^2	-1	-1	-1	-1	1	1	1	1

The first group of columns refers to the levels that the categorical variables x_1 and x_2 take. Thus each row in the contrast matrix represents a possible pairing of x_1 and x_2 values. The second group of columns refers to the values that the effect-coded variables would take, and is formed by applying the coding rules described previously. The third group of columns refers to the values that the interactions between the effect-coded variables would take, and is formed by taking appropriate products between the effect-coded variables in the second group of columns.

The third group of columns reveals how to recover coefficients on interacted effect-coded variables. Consider the first three rows: 1) the coefficient on the interaction $l_1^1 \times l_1^2$ is given by the coefficient on $x_1^1 \times x_1^2$, 2) the coefficient on the interaction $l_2^1 \times l_1^2$ is given by the coefficient on $x_2^1 \times x_1^2$, 3) the coefficient on the interaction $l_3^1 \times l_1^2$ is given by the negative sum of the aforementioned coefficients, $-(b_{11}^{12} + b_{21}^{12})$.

Examining the remaining rows in a similar pattern, focusing particularly on interactions involving omitted levels, reveals the following pattern: to recover the coefficient on an interaction term where both variables are at their base level, take the positive sum of all the coefficients on the interaction terms where neither variable is at their base level. To recover the coefficient on an interaction term where one variable is at its base level and the other is not, take the negative sum of the coefficients on the interaction terms where the former variable is not at its base level and the latter variable is at the same level. For interaction terms where neither variable is at its base level, recovery is not necessary, as the coefficients are directly estimated.

6.D Vuong test for non-nested hypotheses

The Vuong test (Vuong, 1989) is a test for which of two models is closer to the true specification. It is based on distribution theory for the likelihood ratio statistic, and therefore applies to any model estimated by maximum likelihood. The null hypothesis is that the two models are equally close to the true specification, that is:

$$H_0 : \mathrm{E}_0 \left[\ln \frac{f(y|x;\theta)}{g(y|x,\gamma)} \right]$$

where $f(y|x;\theta)$ is the density for one model and $g(y|x;\gamma)$ is the density for the other model. The expectation in the null hypothesis is unknown, but can be approximated by its sample analogue, the likelihood ratio:

$$\mathrm{LR}(\hat{\theta}, \hat{\gamma}) \equiv \ln L_f(\hat{\theta}) - \ln L_g(\hat{\gamma}) = \sum_{i=1}^{N} \ln \frac{f(y_i|x_i,\hat{\theta})}{g(y_i|x_i,\hat{\gamma})}$$

For strictly non-nested models, the test statistic based on the likelihood ratio has a standard normal distribution under the null:

$$\frac{\mathrm{LR}(\hat{\theta}, \hat{\gamma})}{\sqrt{n}\hat{\omega}} \xrightarrow{i.d.} N(0,1)$$

where

$$\hat{\omega}^2 \equiv \frac{1}{N} \sum_{i=1}^{N} \left(\ln \frac{f(y_i|x_i,\hat{\theta})}{g(y_i|x_i,\hat{\gamma})} \right)^2 - \left(\frac{1}{N} \sum_{i=1}^{N} \ln \frac{f(y_i|x_i,\hat{\theta})}{g(y_i|x_i,\hat{\gamma})} \right)^2$$

There are two alternative hypotheses to the null:

$$H_f : \mathrm{E}_0[\ln(f/g)] > 0$$
$$H_g : \mathrm{E}_0[\ln(f/g)] < 0$$

In other words, H_f implies that model F_θ is closer to the true specification than model G_γ, and H_g implies that model G_γ is closer to the true specification than model F_θ.

For a critical value of c, the decision rule is as follows:

- If the test statistic is greater than c, then reject H_0 in favour of H_f.
- If the test statistic is less than $-c$, then reject H_0 in favour of H_g.
- Otherwise, the test statistic is between c and $-c$, and the test cannot discriminate between the two models.

6.E Measuring prediction accuracy using K-fold cross validation

Cross validation is a standard method in the statistics and data mining literatures to estimate out-of-sample prediction error (Hastie et al., 2009). Under cross-validation, the dataset is partitioned into K 'folds'. For each fold, the model of interest is estimated using data excluding that fold, and then validated on that fold. In other words, each fold takes a turn acting as the hold-out (*i.e.*test set). The average error across all the folds is an estimate of the out-of-sample prediction error. Cross validation can be performed on every model specification of interest, yielding a measure of prediction accuracy which can be used to select between specifications. The selected model can then be estimated on the full dataset.

The number of folds, K, is a user-specified parameter which is subject to the bias-variance trade-off. The greater the number of folds, the less the bias, but the greater the variance. The smaller the number of folds, the greater the bias. There is no 'correct' value for the number of folds, but $K = 5$ and $K = 10$ can often be seen in the literature. For this study, $K = 10$ is used, although other values K were also tried. Qualitatively, the results were very similar to $K = 10$. The detailed results are available from the author upon request.

The folds are created by randomly dividing the dataset into portions of nearly equal sizes (that is, if the sample size is not divisible by the number of folds, then the number of observations in each fold may be different by one observation[1]). Because the division is random, standard errors for the estimated prediction error can be obtained by bootstrapping the division. In this study, $R = 250$ bootstrap replications are used.

In addition to the number of folds, a loss function must also be specified. The loss function defines how the error is computed for a given choice probability prediction. In this study, the multinomial deviance loss function recommended by Hastie et al. (2009) is used. It is defined as:

$$L(y, \hat{p}(x)) = -\sum_{j=1}^{J} y_j \ln \hat{p}_j(x)$$

where y_j is a dummy variable indicating whether alternative j was chosen by the respondent and $\hat{p}_j(x)$ is the predicted choice probability for alternative j. For convenience, respondent subscripts i have been suppressed.

6.F Coefficient estimates for attendance scale length models

Table 6.2: (continued on next page)

	1 | 23456	1 | 234 | 56	1 | 2 | 3 | 4 | 5 | 6
ASC.SQ	0.03	0.21***	0.33***
	(0.03)	(0.03)	(0.03)
Mag.0	0.31***	0.48***	0.60***
	(0.05)	(0.05)	(0.05)
amount::1	−0.19***	−0.26***	−0.30***
	(0.01)	(0.02)	(0.02)
amount::23456	0.19***		
	(0.01)		
location:workplace:1	0.19***	0.31**	0.28
	(0.05)	(0.10)	(0.15)
location:community:1	−0.12*	−0.21*	−0.16
	(0.05)	(0.10)	(0.16)
location:church:1	0.00	−0.04	−0.27
	(0.05)	(0.09)	(0.16)

[1]Note that for this discrete choice dataset, each 'observation' is defined as a choice task. For data in long form, one observation therefore corresponds to three rows, because there is one row for each of the three alternatives in a choice task.

Table 6.2: (continued from previous page)

	1 | 23456	1 | 234 | 56	1 | 2 | 3 | 4 | 5 | 6
location:clinic:1	−0.08	−0.06	0.15
	(0.05)	(0.10)	(0.15)
location:workplace:23456	−0.19***		
	(0.05)		
location:community:23456	0.12*		
	(0.05)		
location:church:23456	−0.00		
	(0.05)		
location:clinic:23456	0.08		
	(0.05)		
form:gym:1	−0.07	−0.11	−0.06
	(0.05)	(0.10)	(0.19)
form:medical:1	0.08	0.13	0.21
	(0.05)	(0.10)	(0.19)
form:debit:1	0.09	0.17	0.16
	(0.05)	(0.10)	(0.17)
form:cash:1	−0.11	−0.19	−0.31
	(0.06)	(0.10)	(0.20)
form:gym:23456	0.07		
	(0.05)		
form:medical:23456	−0.08		
	(0.05)		
form:debit:23456	−0.09		
	(0.05)		
form:cash:23456	0.11		
	(0.06)		
condition:weight:1	−0.05	−0.00	0.11
	(0.06)	(0.09)	(0.12)
condition:compliance:1	−0.03	−0.16	−0.27*
	(0.05)	(0.09)	(0.12)
condition:att.comp:1	0.01	0.10	0.15
	(0.05)	(0.09)	(0.11)
condition:attendance:1	0.07	0.07	0.02
	(0.05)	(0.09)	(0.12)
condition:weight:23456	0.05		
	(0.06)		

Table 6.2: (continued from previous page)

	1 | 23456	1 | 234 | 56	1 | 2 | 3 | 4 | 5 | 6
condition:compliance:23456	0.03		
	(0.05)		
condition:att.comp:23456	−0.01		
	(0.05)		
condition:attendance:23456	−0.07		
	(0.05)		
frequency:weekly:1	0.07	0.11	0.16
	(0.05)	(0.09)	(0.13)
frequency:monthly:1	0.10	0.17	0.09
	(0.05)	(0.09)	(0.13)
frequency:quarterly:1	−0.14**	−0.22*	−0.15
	(0.05)	(0.09)	(0.13)
frequency:once:1	−0.03	−0.06	−0.10
	(0.05)	(0.09)	(0.14)
frequency:weekly:23456	−0.07		
	(0.05)		
frequency:monthly:23456	−0.10		
	(0.05)		
frequency:quarterly:23456	0.14**		
	(0.05)		
frequency:once:23456	0.03		
	(0.05)		
amount::234		0.07***	
		(0.01)	
amount::56		0.19***	
		(0.01)	
location:workplace:234		−0.15	
		(0.09)	
location:community:234		0.12	
		(0.09)	
location:church:234		0.01	
		(0.08)	
location:clinic:234		0.02	
		(0.09)	
location:workplace:56		−0.17**	
		(0.06)	

Table 6.2: (continued from previous page)

	1 | 23456	1 | 234 | 56	1 | 2 | 3 | 4 | 5 | 6
location:community:56		0.09	
		(0.06)	
location:church:56		0.03	
		(0.06)	
location:clinic:56		0.04	
		(0.06)	
form:gym:234		0.07	
		(0.08)	
form:medical:234		0.06	
		(0.08)	
form:debit:234		-0.15^{*}	
		(0.08)	
form:cash:234		0.03	
		(0.09)	
form:gym:56		0.04	
		(0.07)	
form:medical:56		-0.19^{**}	
		(0.07)	
form:debit:56		-0.01	
		(0.06)	
form:cash:56		0.16^{*}	
		(0.07)	
condition:weight:234		-0.02	
		(0.07)	
condition:compliance:234		-0.06	
		(0.07)	
condition:att.comp:234		0.12	
		(0.06)	
condition:attendance:234		-0.04	
		(0.07)	
condition:weight:56		0.02	
		(0.08)	
condition:compliance:56		0.22^{**}	
		(0.08)	
condition:att.comp:56		-0.22^{**}	
		(0.08)	

Table 6.2: (continued from previous page)

	1 | 23456	1 | 234 | 56	1 | 2 | 3 | 4 | 5 | 6
condition:attendance:56		−0.03	
		(0.08)	
frequency:weekly:234		0.05	
		(0.07)	
frequency:monthly:234		0.06	
		(0.07)	
frequency:quarterly:234		−0.01	
		(0.07)	
frequency:once:234		−0.10	
		(0.07)	
frequency:weekly:56		−0.16*	
		(0.07)	
frequency:monthly:56		−0.22**	
		(0.07)	
frequency:quarterly:56		0.23**	
		(0.07)	
frequency:once:56		0.16*	
		(0.07)	
amount::2			−0.06*
			(0.02)
amount::3			0.02
			(0.02)
amount::4			0.06**
			(0.02)
amount::5			0.14***
			(0.02)
amount::6			0.14***
			(0.02)
location:workplace:2			0.06
			(0.18)
location:community:2			0.02
			(0.19)
location:church:2			0.03
			(0.18)
location:clinic:2			−0.11
			(0.18)

Table 6.2: (continued from previous page)

	1 | 23456	1 | 234 | 56	1 | 2 | 3 | 4 | 5 | 6
location:workplace:3			−0.01
			(0.16)
location:community:3			0.03
			(0.16)
location:church:3			0.22
			(0.15)
location:clinic:3			−0.24
			(0.17)
location:workplace:4			−0.07
			(0.13)
location:community:4			0.04
			(0.13)
location:church:4			−0.06
			(0.12)
location:clinic:4			0.09
			(0.13)
location:workplace:5			−0.09
			(0.10)
location:community:5			−0.08
			(0.10)
location:church:5			0.08
			(0.10)
location:clinic:5			0.08
			(0.10)
location:workplace:6			−0.18*
			(0.08)
location:community:6			0.15*
			(0.08)
location:church:6			−0.00
			(0.07)
location:clinic:6			0.03
			(0.07)
form:gym:2			0.01
			(0.21)
form:medical:2			0.24
			(0.21)

Table 6.2: (continued from previous page)

	1 | 23456	1 | 234 | 56	1 | 2 | 3 | 4 | 5 | 6
form:debit:2			0.02
			(0.19)
form:cash:2			−0.27
			(0.22)
form:gym:3			−0.04
			(0.15)
form:medical:3			−0.12
			(0.15)
form:debit:3			−0.08
			(0.14)
form:cash:3			0.23
			(0.16)
form:gym:4			0.05
			(0.11)
form:medical:4			0.04
			(0.11)
form:debit:4			−0.10
			(0.11)
form:cash:4			0.02
			(0.11)
form:gym:5			0.08
			(0.10)
form:medical:5			−0.14
			(0.10)
form:debit:5			−0.10
			(0.09)
form:cash:5			0.16
			(0.10)
form:gym:6			−0.05
			(0.10)
form:medical:6			−0.23*
			(0.10)
form:debit:6			0.10
			(0.10)
form:cash:6			0.17
			(0.10)

Table 6.2: (continued from previous page)

	1 | 23456	1 | 234 | 56	1 | 2 | 3 | 4 | 5 | 6
condition:weight:2			−0.26
			(0.14)
condition:compliance:2			0.15
			(0.13)
condition:att.comp:2			0.17
			(0.13)
condition:attendance:2			−0.06
			(0.14)
condition:weight:3			0.04
			(0.13)
condition:compliance:3			−0.10
			(0.11)
condition:att.comp:3			−0.06
			(0.11)
condition:attendance:3			0.13
			(0.11)
condition:weight:4			0.04
			(0.12)
condition:compliance:4			−0.15
			(0.11)
condition:att.comp:4			0.19
			(0.10)
condition:attendance:4			−0.09
			(0.11)
condition:weight:5			0.01
			(0.13)
condition:compliance:5			0.31**
			(0.12)
condition:att.comp:5			−0.34**
			(0.12)
condition:attendance:5			0.01
			(0.12)
condition:weight:6			0.06
			(0.13)
condition:compliance:6			0.06
			(0.13)

Table 6.2: (continued from previous page)

	1 | 23456	1 | 234 | 56	1 | 2 | 3 | 4 | 5 | 6
condition:att.comp:6			−0.11
			(0.12)
condition:attendance:6			−0.01
			(0.12)
frequency:weekly:2			0.11
			(0.15)
frequency:monthly:2			0.25
			(0.14)
frequency:quarterly:2			−0.19
			(0.14)
frequency:once:2			−0.17
			(0.15)
frequency:weekly:3			0.06
			(0.14)
frequency:monthly:3			0.06
			(0.12)
frequency:quarterly:3			−0.18
			(0.13)
frequency:once:3			0.06
			(0.13)
frequency:weekly:4			0.02
			(0.11)
frequency:monthly:4			0.01
			(0.10)
frequency:quarterly:4			0.08
			(0.11)
frequency:once:4			−0.10
			(0.11)
frequency:weekly:5			−0.09
			(0.11)
frequency:monthly:5			−0.07
			(0.11)
frequency:quarterly:5			0.17
			(0.11)
frequency:once:5			−0.01
			(0.12)

Table 6.2: (continued from previous page)

	1 | 23456	1 | 234 | 56	1 | 2 | 3 | 4 | 5 | 6
frequency:weekly:6			-0.26^{*}
			(0.11)
frequency:monthly:6			-0.34^{**}
			(0.10)
frequency:quarterly:6			0.28^{**}
			(0.11)
frequency:once:6			0.32^{**}
			(0.11)
Log-likelihood	−2410.95	−2389.88	−2379.63
AIC	4851.90	4835.77	4893.25
BIC	4939.27	4998.86	5283.52
$\bar{\rho}^2$	0.10	0.10	0.09
N	2502.00	2502.00	2502.00

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 6.2: Coefficient estimates for models under selected dichotomous, trichotomous and original six-point attendance scales. Reported coefficients include recovered effects (see section 6.C) and standard errors approximated by the delta method. Interaction terms are reported as `attribute:attribute level:attendance level`, where the attendance level may be defined by multiple points from the original six-point attendance scale.

6.G Table of coefficients for latent class logit model

Table 6.3: (continued on next page)

	LCL
theta1	$-1.76\ (0.23)^{***}$
theta2	$-1.82\ (0.21)^{***}$
theta3	$-0.90\ (0.18)^{***}$
theta4	$-0.66\ (0.10)^{***}$

Table 6.3: (continued from previous page)

	LCL
ASC.SQ	−0.05 (0.10)
Mag.0	0.43 (0.11)***
condition1	0.11 (0.05)*
condition2	−0.08 (0.05)
condition3	−0.07 (0.04)
frequency1	0.11 (0.05)*
frequency2	0.03 (0.05)
frequency3	0.03 (0.05)
form1	−0.15 (0.05)**
form2	−0.16 (0.05)***
form3	0.12 (0.05)*
1.log(amount + 1)	0.28 (0.13)*
1.location1	−5.81 (1.28)***
1.location2	1.91 (0.47)***
1.location3	0.21 (0.56)
2.log(amount + 1)	−0.34 (0.08)***
2.location1	4.95 (0.86)***
2.location2	4.88 (0.80)***
2.location3	−14.68 (2.40)***
3.log(amount + 1)	2.57 (0.50)***
3.location1	0.23 (0.23)
3.location2	0.20 (0.24)
3.location3	−0.59 (0.28)*
4.log(amount + 1)	−0.40 (0.05)***
4.location1	−0.01 (0.14)
4.location2	0.11 (0.14)
4.location3	−0.31 (0.17)
5.log(amount + 1)	0.28 (0.05)***
5.location1	−0.15 (0.08)*
5.location2	−0.02 (0.07)
5.location3	0.16 (0.06)**
Log-likelihood	−4387.84

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

6.H Table of coefficients for triangular MXL

	Triangular
ASC.SQ	−0.09 (0.09)
Mag.0	0.11 (0.10)
condition1	0.07 (0.04)
condition2	−0.05 (0.04)
condition3	−0.11 (0.04)**
frequency1	0.13 (0.05)**
frequency2	−0.01 (0.04)
frequency3	0.04 (0.04)
form1	−0.13 (0.04)**
form2	−0.15 (0.04)***
form3	0.12 (0.04)**
location1.c	−0.20 (0.12)
location1.s1	−0.00 (0.11)
location1.s2	0.02 (0.29)
location2.c	0.14 (0.04)***
location2.s1	0.00 (0.17)
location2.s2	−0.00 (0.17)
location3.c	−0.27 (−)
location3.s1	0.01 (−)
location3.s2	−0.31 (−)
log(amount + 1).c	0.48 (0.15)**
log(amount + 1).s1	1.45 (0.17)***
log(amount + 1).s2	0.88 (0.32)**
Log-likelihood	−4481.91

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

6.I Table of coefficients for Johnson SB MXL

	Johnson SB
ASC.SQ	-0.15 (0.09)
Mag.0	0.05 (0.10)
condition1	0.07 (0.04)
condition2	-0.04 (0.04)
condition3	-0.10 (0.04)*
frequency1	0.13 (0.05)**
frequency2	-0.01 (0.04)
frequency3	0.05 (0.04)
form1	-0.13 (0.04)**
form2	-0.16 (0.04)***
form3	0.14 (0.04)**
location1.gamma	0.86 (0.11)***
location1.delta	-0.00 ($-$)
location1.xi	-0.23 (0.06)***
location1.lambda	-0.19 (0.27)
location2.gamma	-1.41 (1.43)
location2.delta	-8.75 ($-$)
location2.xi	0.12 (0.38)
location2.lambda	0.04 (0.69)
location3.gamma	2.67 (0.33)***
location3.delta	0.20 (0.11)
location3.xi	-0.24 (0.05)***
location3.lambda	46.35 ($-$)
log(amount + 1).gamma	-4.39 (0.86)***
log(amount + 1).delta	2.06 (0.82)*
log(amount + 1).xi	-8.64 (1.41)***
log(amount + 1).lambda	10.08 (1.60)***
Log-likelihood	-4472.65

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Bibliography

Bech, Mickael and Gyrd-Hansen, Dorte. Effects coding in discrete choice experiments. *Health Economics*, 14(10):1079–1083, 2005. ISSN 1099-1050. doi: 10.1002/hec.984. URL http://dx.doi.org/10.1002/hec.984.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer, 2009. ISBN 9780387848587.

Holmes, Thomas P. and Adamowicz, Wiktor L. *A Primer on Nonmarket Valuation*, chapter Attribute-Based Methods, pages 171–219. Kluwer Academic Publishers, Dordrecht, 2003.

Hoyos, David. The state of the art of environmental valuation with discrete choice experiments. *Ecological Economics*, 69(8):1595 – 1603, 2010. ISSN 0921-8009. doi: http://dx.doi.org/10.1016/j.ecolecon.2010.04.011. URL http://www.sciencedirect.com/science/article/pii/S0921800910001540.

Lancsar, Emily and Louviere, Jordan. Conducting discrete choice experiments to inform healthcare decision making. *PharmacoEconomics*, 26(8):661 – 677, 2008. ISSN 11707690. URL http://link.springer.com/article/10.2165/00019053-200826080-00004.

Vuong, Quang H. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):pp. 307–333, 1989. ISSN 00129682. URL http://www.jstor.org/stable/1912557.